Characterization of Cancer Genomes through Systematic Analyses of Oncogenomic Data Assemblies

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

HAOYANG CAI

aus China

Promotionskomitee

Prof. Dr. Christian von Mering (Vorsitz der Dissertation)

Dr. Michael Baudis (Leitung der Dissertation)

Prof. Dr. Mark D. Robinson

Prof. Dr. Homayoun Bagheri

Prof. Dr. Nuria Lopez-Bigas

Zürich, 2013

ACKNOWLEDGEMENTS

I dedicate this thesis to my parents who always and unconditionally supported me in finding and pursuing my path.

I would like to express my deepest appreciation to my supervisor Michael Baudis for giving me the opportunity to work in this exciting field. I learnt a lot in the process, and Michael, thank you for guiding me to take the giant leap forward from computer science to life science.

Very special thanks I give to Christian von Mering, Mark Robinson, Homayoun Bagheri and Nuria Lopez-Bigas, whose stimulating suggestions and encouragement helped me to make progress with my projects. Without your continuous help and support I would never finish my Ph.D.

I also wish to thank Nitin Kumar, Ni Ai and Saumya Gupta for making the lab a friendly working place, and for the happy times we shared. Many thanks go to Julia Jaeger, Juan Carrillo, Susanna Bachmann, Bettina Rausch-Malina, Eveline Bergmüller, Oleg Georgiev and Werner Wolz. Your patience and help are greatly appreciated.

Last but not least, I would like to thank my good friends and colleagues, Xue Zheng, Mingcong Wang, Qiutan Yang, Sheng Zeng, Xiaobei Zhou, von Mering group and Robinson group for encouraging me to pursue my research, for providing invaluable assistance, and for motivating me to continue on my path even when it was difficult.

Thank you all for making my Ph.D. an enjoyable and memorable time!

ABSTRACT

Cancer is the most common genetic disease in humans. It has been estimated that more than 10 million new cancer patients are detected worldwide each year. In the last decades, many efforts have been made by the research community to contribute to the fight against cancer. These works greatly expanded our understanding of the disease. However, the exact mechanisms of cancer initiation and progression remain elusive. The research on cancer genomes has focused on the identification of DNA sequence mutations and chromosomal rearrangements. Some of these somatic alterations can confer a growth advantage to cancer cells and promote cancer development. Mutated genes in cancer genomes can be potential new drug targets or serve as biomarkers for the improvement of diagnostics and therapy.

Today, high-throughput genome-wide profiling technologies allow us to characterize the molecular profiles of cancer samples on various levels, including copy number alterations, gene expression, point mutations and epigenetic marks. Cancer research has gradually shifted from single experiments to large-scale “omics” data analysis approaches. It is an exciting, but challenging work. Our group aims to develop reliable and robust methods to characterize cancer genomes by analyzing large-scale oncogenomic datasets.

During the last 4 years, I have focused my efforts on using systems biology and statistical methods to model and annotate genomic array data in human cancer. My research is based on a data collection and re-analysis project that generates very large amounts of microarray data. Computational biology approaches were applied on this dataset for data mining. We collected more than 40000 arrays, including comparative genomic hybridization (CGH) and SNP (single nucleotide polymorphism) arrays, from several public databases. A pipeline was developed to process raw data and determine copy number aberrations (CNAs). All data was converted to a unified and structured format, and stored in our arrayMap database, together with available clinical information. We also set up an online website for providing this resource to the research community.

Based on the large-scale CNA data in our database, the second project aimed to explore the correlation between CNAs and local gene density across cancer genomes. Through a genome binning method, I found that focal CNAs are significantly enriched in gene-rich regions. In addition, this positive correlation is not only driven by cancer genes. Since this result is derived from more than 16000 cancer samples, it provides a global insight into the relationship between cancer genome instability and structure from a new perspective. The enrichment reveals that there may be a non-neutral selection pressure for CNA regions across the genome. The observed significant positive correlation in this project may enable a better elucidation of mechanisms by which CNAs contribute to tumor development, and promote a more systematic understanding of cancer.

The third project presented here is related to a new phenomenon, termed “chromothripsis”, found in cancer development. In this type of events, contiguous chromosomal regions are fragmented into many pieces, and the cell’s DNA repair machinery randomly fuses these segments together to rescue the genome. This is quite different from the classical step-by-step model of cancer development. We developed an algorithm based on scan statistics to automatically detect chromothripsis-like patterns, and identify both size and location of the involved regions. From our input of 22,347 high quality arrays, we identified 918 chromothripsis cases, representing 132 cancer types. The results from this dataset provide several new insights regarding the distribution of chromothripsis-like patterns and a comprehensive estimation of chromothripsis incidence in a large range of cancer entities. Importantly, our work partly overcomes the limitation of individual research projects resulting from the relatively low incidence of chromothripsis in cancer samples available. An investigation into the affected chromosomal regions supports breakage-fusion-bridge cycles as one of the potential underlying mechanisms. Finally, we evaluated the clinical associations of chromothripsis and found that this event may be associated with a poor outcome. The observed chromothripsis events in our project may reflect on heterogenous biological phenomena, and probably vary in their specific impact on oncogenesis.

Taken together, the results presented in this thesis characterize the cancer genome by large-scale oncogenomic array data, and further elucidate the potential mechanisms underlying cancer development.

ZUSAMMENFASSUNG

Krebs ist die häufigste auf genomische Veränderungen zurückzuführende Erkrankung. Weltweit wird in jedem Jahr bei geschätzt 10 Millionen Menschen eine Krebserkrankung diagnostiziert. In den letzten Jahrzehnten wurden grosse Anstrengungen unternommen, um Krebs unter Einsatz wissenschaftlicher Methoden zu bekämpfen. Obwohl dies unser Verständnis bösartiger Erkrankungen massiv erweitert hat, sind trotzdem viele Mechanismen der Krebsentstehung und Ausbreitung noch immer unvollständig charakterisiert.

Bisher wurde im Genom von Krebszellen hauptsächlich nach Mutationen in der DNA-Sequenz sowie nach chromosomalen Veränderungen gesucht. Einige dieser Veränderungen führen bekanntermassen zu einem Wachstumsvorteil von Krebszellen und begünstigen die Krebsentwicklung. Die am häufigsten mutierten Gene in Krebszellen stellen potenzielle Angriffspunkte für neue Medikamente dar oder könnten als Biomarker zur verbesserten Diagnose und Therapie dienen.

Heutzutage können die molekularen Profile von Krebszellproben auf unterschiedlichen Ebenen durch genomweite Hochdurchsatztechnologien untersucht werden – von chromosomalen Aberrationen und Punktmutationen, über Modulation der Genexpression bis hin zu epigenetischen Veränderungen. Die molekulare Krebsforschung hat sich dabei von der Beurteilung einzelner Experimente hin zur umfassenden “omics”-Analyse entwickelt. Diese Untersuchungen sind sowohl hochinteressant von ihrer Thematik, als auch anspruchsvoll in der Durchführung. Unsere Forschungsgruppe beteiligt sich hier durch die Entwicklung von verlässlichen und robusten Methoden zur umfassenden Analyse von Krebsgenomen.

Während der letzten vier Jahre habe ich mich hauptsächlich mit systematischer und statistischer Analyse zur Modellierung und Annotation genomischer Array-Daten von menschlichen Krebszellen beschäftigt. Meine Forschung basiert auf einem Datenkollektions-und Wiederanalyse-Projekt, das eine grosse Menge Microarray-Daten

produziert hat. Zum “data mining” wurden computer-gestützte Methoden auf diesen Datensatz angewendet. Wir haben hierzu Daten von mehr als 40'000 Genomanalysen aus öffentlich zugänglichen Datenbanken gesammelt, einschliesslich Daten von komparativer Genomhybridisierung (CGH) und Einzelnukleotidvariationen (SNP). Ich entwickelte einen definierten Arbeitsablauf um Rohdaten zu prozessieren und Veränderungen der regionalen DNA-Kopienanzahl (CNAs) zu analysieren. Alle Daten wurden in ein einheitlich strukturiertes Format gebracht und zusammen mit verfügbaren klinischen Informationen in unserer “arrayMap”-Datenbank gespeichert. Zusätzlich haben wir eine Online-Resource entwickelt, um diese Daten der Forschungsgemeinschaft zugänglich zu machen.

Basierend auf der umfassenden CNA-Datensammlung in unserer Datenbank wurde in einem zweiten Projekt der Zusammenhang zwischen CNAs und lokaler Gendichte analysiert. Durch “genome binning” habe ich herausgefunden, dass fokale CNAs in genreichen Regionen signifikant angereichert sind. Zusätzlich konnte ich zeigen, dass diese Korrelation nicht nur durch Krebsgene verursacht wird. Da diese Erkenntnis auf mehr als 16'000 Proben basiert, erlaubt sie einen globaleren Einblick ins Verhältnis zwischen Instabilität und Struktur von Krebsgenomen. Diese Anreicherung zeigt, dass ein nicht-neutraler Selektionsdruck für CNA-Regionen über das ganze Genom bestehen könnte. Die in diesem Projekt beobachtete signifikant-positive Korrelation könnte einen besseren Einblick in die Mechanismen ermöglichen, durch die CNAs zur Tumorentwicklung beitragen und für ein systematischeres Verständnis von Krebs sorgen.

Im dritten Projekt wurde ein neues Phänomen namens Chromothripsis untersucht, das kürzlich für den Prozess der Krebsentwicklung beschrieben wurde. Hierbei werden zusammenhängende chromosomale Regionen in viele Teile fragmentiert, worauf die zelluläre DNA-Reparaturmaschinerie diese Stücke zur Rettung des Genoms zufällig wieder zusammensetzt. Dies stellt einen fundamentalen Unterschied zum klassischen schrittweisen Modell der Krebsentstehung dar. Unter Verwendung einer “Scan-Statistcs” basierten Methodik haben wir einen Algorithmus entwickelt, der automatisch Chromothripsis-artige Muster detektiert und sowohl Grösse als auch Lokalisation der beteiligten Regionen identifiziert. In den 22'347 untersuchten Array-datensätzen konnten wir 918 Fälle von Chromothripsis -artigen Mustern aus 132 unterschiedlichen Krebsarten beschreiben. Die Ergebnisse dieser Analyse erlauben neue Einblicke in die Verteilung von Chromothripsis-artigen Mustern und eine umfassende Einschätzung der Chromothripsis-Häufigkeit in einer grossen Anzahl von Krebsfällen. Unsere Arbeit konnte dabei teilweise

die Grenzen überwinden, die sich aus der relativ kleinen Anzahl der verfügbaren Chromothripsis-Fälle in individuellen Forschungsprojekten ergeben. Die Analyse der betroffenen Chromosomenregionen unterstützt “breakage-fusion-bridge-cycles” als einen der möglicherweise zugrunde liegenden Mechanismen. Untersuchungen der Assoziatione klinischer Daten mit Chromothripsis-artigen Genommustern und zeigte ein häufiges Zusammenfallen dieser Ereignisse mit einer schlechten klinischen Prognose. Aus unseren Daten ergeben sich aber auch Hinweise, dass die beobachteten Chromothripsis-artigen Fälle möglicherweise ein heterogenes biologisches Phänomen darstellen, und damit in ihrem spezifischen Einfluss auf die Onkogenese variieren könnten.

Zusammenfassend charakterisieren die in dieser Arbeit präsentierten Resultate das Krebsgenom durch umfassende Analysen onkogenomische Array-Daten und bringen neue Einblicke der Krebsentwicklung potenziell zugrunde liegende Mechanismen.

PUBLICATIONS

PUBLICATIONS INCLUDED IN THIS THESIS:

Paper I

Cai H, Kumar N, Baudis M
arrayMap: a reference resource for genomic copy number imbalances in human malignancies.
PLoS One. 2012;7(5):e36944.

Paper II

Cai H, Kumar N, Ai N, von Mering C, Baudis M
The landscape of cancer genomes reveals correlation between somatic copy number aberrations and fundamental genome structure.

Paper III

Cai H, Kumar N, Bagheri HC, von Mering C, Robinson MD, and Baudis M
Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genomes.

Paper IV

Kumar N, Cai H, von Mering C, Baudis M
Specific genomic regions are differentially affected by copy number alterations across distinct cancer types, in aggregated cytogenetic data.
PLoS One. 2012;7(8):e43689.

Paper V

von Bueren AO, Gerss J, Hagel C, Cai H, Remke M, Hasselblatt M, Feuerstein BG, Pernet S, Delattre O, Korshunov A, Rutkowski S, Pfister SM, Baudis M
DNA copy number alterations in central primitive neuroectodermal tumors and tumors of the pineal region: an international individual patient data meta-analysis.
J Neurooncol. 2012 Sep;109(2):415-23.

CONTENTS

ACRONYMS

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
AR	Allelic Ratio
BAC	Bacterial Artificial Chromosome
BIR	Break-induced Replication
BISRS	Break-induced Serial Replication Slippage
CBS	Circular Binary Segmentation
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CNA	Copy Number Aberration
DNA	Deoxyribonucleic Acid
FISH	Fluorescent in situ Hybridization
FoSTeS	Fork Stalling and Template Switching
GISTIC	Genomic Identification of Significant Targets in Cancer
GO	Gene Ontology
ICD-O	International Classification of Diseases for Oncology
LFS	Li-Fraumeni Syndrome
LOH	Loss of Heterozygosity
MMBIR	Microhomology-mediated Break-induced Replication
MMRDR	Microhomology-mediated Replication-dependent Recombination
NAHR	Non-allelic Homologous Recombination
NGS	Next-generation Sequencing
NHEJ	Non-homologous End Joining
NUSE	Normalized Unscaled Standard Error
RLE	Relative Log Expression
RMA	Robust Multichip Analysis
ROC	Receiver Operating Characteristic
ROMA	Representational Oligonucleotide Microarray Analysis
SNP	Single Nucleotide Polymorphism
SSA	Single-strand Annealing
WHO	World Health Organization
YAC	Yeast Artificial Chromosome

Part I

INTRODUCTION AND CONTEXT

1 INTRODUCTION

Cancer is a class of diseases that are responsible for the overall leading cause of death worldwide[1]. There are over 200 different types of cancer, although this number depends on the various classification systems. The overall defining feature of cancer is an out-of-control cell growth that may form tumors and harm the body[2,3]. Tumors with un-invasive growth and without metastasis are considered to be benign. Malignant tumors can spread to other organs in the body through the blood and lymphatic systems, and finally cause the death of cancer patients[4].

The field of cancer research had begun by the end of the 19th century. David von Hansemann[5] and Theodor Boveri[6] first reported the role of the "genome" in cancer development at the turn of the 20th century. They observed the aberration of chromosomes under the microscope, and proposed the hypothesis that cancers are abnormal clones of cells, and caused by change of hereditary material. Almost 50 years later, the discovery of DNA as the hereditary material was made[7]. Furthermore, Watson and Crick described the structure of DNA[8]. Thus, the idea that damage DNA and genome mutations cause cancer was accepted.

The next spectacular success in cancer research came from the discovery of Philadelphia (Ph) chromosome, which is a karyotypic marker in chronic myeloid leukemia[9,10]. This chromosome is caused by a translocation between chromosome 9 and chromosome 22. At the beginning, the translocation occurs in a single bone marrow cell. After the process of clonal expansion, it finally gives rise to the leukemia. This was the first consistent chromosome alteration in human cancer and provided an evidence for chromosomal abnormality as the direct cause of neoplasia[9,10].

The discovery of Philadelphia chromosome greatly stimulated interest in cancer research. At the same time, new technologies had been developed to identify genetic alterations in whole cancer genomes, and provided new insights into the mechanisms of human tumors at the molecular level. For example, comparative genomic hybridization (CGH) is a molecular cytogenetic method for genome-wide screening of DNA copy number

aberrations[11,12,13]. With the successive developments of cancer research tools, focuses of studies have expanded from single genes to pathways, and from classical genetics to systems biology. It is clear that, in the future, systematic functional approaches will be employed to integrate genome-scale information, and obtain a comprehensive molecular understanding of cancer. In clinical practice, the products of decades of research have greatly contributed to the development of novel diagnostic methods and therapeutic strategies[3,14]. Today, cancer biology has become one of the most exciting fields in life sciences, and continuously provides us with new insights into our most common genetic disease.

The work presented in this thesis represents an effort to analyze large-scale data sets derived mainly from array comparative genomic hybridization experiments for different cancer types, through bioinformatics and statistical methodology. Our studies focus on elucidating and modeling mechanisms for cancer development. Through the data collection project, we also provide the research community a valuable resource for oncogenomic data. It is an *in silico* systems biology platform for the comprehensive study of cancer.

2 THE CANCER GENOME

2.1 CANCER AND COPY NUMBER ABERRATION

Cancer as a leading cause of disease is responsible for approximately 12% of deaths around the world[1]. Millions of people die from cancer each year. Although there are many types of cancer, they have some common characteristics. Cancer cells grow out of control and form solid masses (tumors) or diffusely invasive cells with expansive growth characteristics (e.g. bone marrow replicate in leukemias)[1]. In the beginning, these cells usually suffer from DNA damage. These damages are neither repaired nor induce cell death. These tumor cells continue to generate new abnormal cells. Malignant tumors may invade, and subsequently destroy, the healthy tissues adjacent to them. The invasion of essential organs, such as the liver, brain, kidneys or lungs, may eventually cause loss of organ function and ultimately death. Moreover, metastasis may occur if cancer cells spread from their original location, i.e. the primary tumor, to a distant organ[15]. A recent review summarized six hallmarks of cancer[2]:

- Sustaining of proliferative signaling
- Evading growth suppressors
- Resisting cell death
- Enabling replicative immortality
- Inducing angiogenesis
- Activating invasion and metastasis

These hallmarks represent six biological capabilities acquired during cancer progression, indicating the complexity of the cancer genome. Treatments modalities in cancer include radiation therapy[16], chemotherapy[17], surgery, immunotherapy[18], targeted agents[19]. Which treatment or combination of treatments should be chosen depends on the type, size and location of cancer, as well as the patient's age and physical condition.

Cancer can affect every organ and cell type in the human body, though some organs are more frequently affected than others. Even in one organ, there are usually many potential

cellular types that cancer can develop. For example, there are several types of cells in the lung, which correspond to different cancer types. If the cancer develops from squamous cells (i.e. bronchial lining), it is called squamous-cell lung carcinoma, that developing from gland cells is lung adenocarcinoma[20,21], that from neuroendocrine cells in the bronchus called small-cell lung carcinoma. In general, more than 200 different types of cancer may arise in the human body. There are many ways for cancer classification, and some examples are listed in Table 1. The commonly used methods based on site of origin or histological types of cancer cells[1]. The treatment will be different depending on the type of cancer. Therefore, the accurate categorization of tumors and subtypes is important for both therapy and prognostic assessment.

When based on the site of origin, cancers can be classified by the affected organs. In this case, the main types of cancer would include brain cancer, breast cancer, lung cancer, liver cancer, kidney cancer, colorectal cancer, etc. Figure 1 shows the proportion of different cancer types based on the GLOBOCAN[22] estimates in 2008. This figure also compares the incidences with ratios of cancers that are derived from our collection of oncogenomic microarrays[23]. The genomic array data collection and processing procedures are presented in Chapter 4.

Table 1. Several cancer classification systems

Classification system	Description	Approximate number of types
ICD-O-3 morphology	Used principally in tumor or cancer registries for coding the histology of neoplasms, usually obtained from a pathology report	873
ICD-O-3 topography	Used principally in tumor or cancer registries for coding the site of neoplasms	98
SEER	Surveillance, Epidemiology and End Results	70
Clinico-pathological entities	Combination of morphology and topography	83

At the moment, the international standard for cancer classification is ICD-O-3 (International Classification of Diseases for Oncology, 3rd Edition)[24]. It is a coding system developed by the World Health Organization (WHO), that integrates tumor site (topography) and the histology (morphology)[24]. The code is composed of five digits ranging from 8000/0 to

9989/3. The first four digits represent the histological term of the neoplasm. The separate one-digit code after the slash is the behavior code, which indicates whether the tumor is malignant, benign, *in situ*, or uncertain. As an example of the ICD-O coding system on the central nervous tumors, Table 2 lists the names and codes of all cranial and paraspinal nerves tumors based on the 2007 WHO classification of tumors. In my following works, I mainly use the ICD-O-3 system to annotate oncogenomic data.

Table 2. Tumors of cranial and paraspinal nerves

Tumors	ICD-O code
Schwannoma (neurilemoma, neurinoma)	9560/0
Cellular	9560/0
Plexiform	9560/0
Melanotic	9560/0
Neurofibroma	9540/0
Plexiform	9550/0
Perineurioma	
Perineurioma, NOS	9571/0
Malignant perineurioma	9571/3
Malignant peripheral nerve sheath tumour (MPNST)	
Epithelioid MPNST	9540/3
MPNST with mesenchymal differentiation	9540/3
Melanotic MPNST	9540/3
MPNST with glandular differentiation	9540/3

Clinical information is particularly important in cancer research. Besides morphology and topography, other clinical data that I collected including patient age, stage, grade, sample source, follow up and information on death. Cancer stage describes the extent to which a cancer has developed by spreading. It is essential in choosing a treatment method and assessing prognosis. Several staging systems are used by the research community to classify tumors. In general, they take into account the size and location of the tumor, and the spread in reference to nearby lymph nodes[1]. The two most commonly used staging methods are TNM and the four-stage rating scale. The TNM system evaluates tumors for size (T), presence or absence of regional lymph node involvement (N), and distant metastasis (M)[25]. Specifically, “T” ranges from 0 to 4, where higher values represent increasing tumor size and involvement. Similarly, “N” ranges from 0 to 3, to indicate the degrees of lymph node involvement, with 0 being the lowest and 3 being the highest degrees. “M” can be 0 or 1, and provides the information if there is evidence of distant spread. The four-stage rating system assigns a number from I to IV to a tumor, with I represents an isolated cancer and IV means the cancer has spread to other organs, i.e. from early stage to advanced stage. In short, stage is one of the important tumor characteristics, and helps in cancer treatment. The two staging systems can be converted

to each other by transformation of parameters, as is shown in Table 3. For the data annotation work in Chapter 4, I used this table to unify cancer stages.

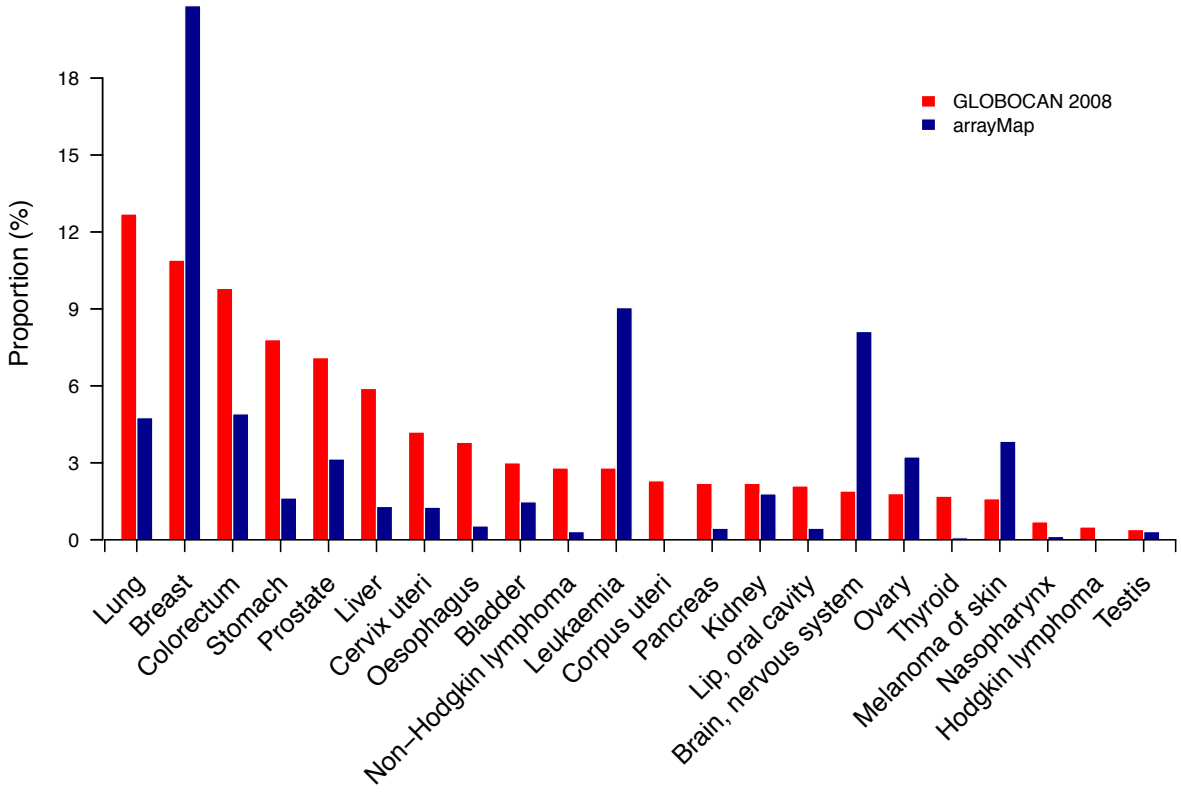


Figure 1. Incidence of different cancer types obtained from GLOBOCAN 2008 and arrayMap. Red and blue bars represent data from GLOBOCAN and arrayMap, respectively. The cancer type data is derived from GLOBOCAN.

Our understanding of the cancer genome has been improved significantly in recent decades, through the rapid advances in molecular biology technology and cytogenetics. These achievements provide a promise for further exploration of the causes of tumor, and allow the development of methods for early detection and cure of cancer. In the last five years, thousands of high resolution genomic or expression microarray experiments have been performed, hundreds of cancer genomes have been sequenced, and even more will follow. These studies generated global profiles of genomic rearrangements, gene expression and somatic mutations in cancer. It will be exciting to perform large-scale data analysis with this wealth of data and results. Actually, numerous cancer genes, candidate biomarkers and therapeutic targets have been discovered. In a more comprehensive level, analyses of pathways, regulatory and functional networks in human cancer have been carried out. In order to systematically store and retrieve these valuable cancer genome

data, many efforts have been devoted to create various repositories, such as GEO[26], ArrayExpress[27], TCGA[28,29] and the UCSC Cancer Genomics Browser[30]. These public databases promote meta-analysis, which will help to elucidate the molecular underpinnings of cancer.

It is now clear that cancer is a genetic disease, caused by a series of acquired or inherited mutations. Cancer patients can inherit abnormal DNA from their parents. However, most DNA damage is caused by replication errors resulting from normal cell proliferation or induced by carcinogenic factors present in our environment. Some human behaviors may lead to higher cancer incidence rates. For example, there is a strong correlation between lung cancer and smoking[31], and between liver cancer and over-consumption of alcohol[32], and also between skin cancer and overexposure to sunlight[33], and so on. In general, common genetic mutations in cancer may be broadly classified as genomic rearrangements and sequence mutations[4]. This thesis focuses on the study of genomic rearrangements in large-scale.

Table 3. Cancer stage grouping

Stage	Primary Tumor (T)	Lymph Nodes (N)	Metastasis (M)
Stage I	T1	N0	M0
Stage II	T1	N1	M0
	T2	N0	M0
	T2	N1	M0
Stage III	T1	N2	M0
	T2	N2	M0
	T3	N0	M0
	T3	N1	M0
	T3	N2	M0
Stage IV	T4	Any N	M0
	Any T	N3	M0
	Any T	Any N	M1

Genomic rearrangements involve changes to a whole chromosome or large chromosomal regions. The related genomic fragments may be constituted by gains, losses, amplifications, insertions, inversions or translocations[34]. Both classical cytogenetic and molecular biological techniques have been employed to study genomic rearrangements. At least three underlying mechanisms of genomic recombination events have been well described by previous publications.

- **Non-homologous end joining (NHEJ)**. NHEJ is the most comprehensive DNA repair mechanism in higher eukaryotes[35]. It involves ligation of any two broken ends independently of DNA homology. Two sub-pathways involved into NHEJ: classical and non-classical.

- **Homologous recombination**. It occurs during the repair of DNA double-strand breaks. DNA sequences which show extensive homology can cause chromosomal mis-pairing[35]. Four pathways are implicated in homologous recombination: non-allelic homologous recombination[35,36] (NAHR), break-induced replication[37] (BIR), single-strand annealing[36,38] (SSA) and gene conversion[37].

- **Microhomology-mediated replication-dependent recombination (MMRDR)**. They are DNA replication-based models and have been used to explain the occurrence of strand misaligning by microhomology[36]. It is a general term, including several models, such as microhomology-mediated break-induced replication[36,39] (MMBIR), break-induced serial replication slippage[40] (BISRS), fork stalling and template switching[41,42] (FoSTes).

Genomic rearrangements can be further subdivided into copy number neutral rearrangements, including inversions and translocations, and copy number alterations, including duplications, deletions and amplifications[43,44]. DNA copy number aberrations are a hallmark of solid tumors[3]. They may localize and disturb oncogenes or tumor suppressors and contribute to cancer progression[45]. Furthermore, genomic imbalances exhibit distinct landscapes in many tumor entities, and have been used for tumor subclassification, and to aid diagnosis and prognosis[46,47,48].

Various types and sizes of genomic imbalances can be discerned. Usually, high-level amplifications can be as small as 100 kb in length, and genomic gains and losses are ranging from small focal regions to the whole chromosome[49]. Focal changes are useful for screening novel cancer gene candidates[50]. The mechanism by which CNAs contribute to cancer development is still poorly understood. One hypothesis is that genomic changes increase or decrease gene expression levels of one or more cancer genes[51,52]. Therefore, integrating CNAs with gene expression data would provide more meaningful results by helping to identify driver mutations. Another hypothesis is correlated with breakpoints, which often lead to the formation of oncogenic fusion genes[53,54].

CNAs can be detected by array CGH technology[55], which is described in detail in the next chapter. By definition, array CGH is not able to detect copy number neutral rearrangements. However, other genome-wide molecular screening methods, such as massively parallel paired-end sequencing, allow the detection of such genomic

alterations[56]. Since CNAs are a valuable resource for cancer research, we launched the arrayMap project to collect and annotate genomic arrays. It provides an overview of copy number abnormalities in human cancer from CGH experiments[23]. The database is continuously updated by extracting data from associated resources. The results of this study are presented in Chapter 4.

2.2 CANCER GENES

Genome sequence mutations, together with DNA copy number alterations, represent the basis for cancer development[4]. These mutations may result in altered cancer genes. Two types of cancer genes are responsible for the initiation and progression of cancers: oncogenes and tumor suppressor genes[45]. In most cases, the presence of a mutation in a single gene is not sufficient in itself to cause cancer, because of the existence of protective mechanisms of the body. Cancer is likely to occur when several genes are defective.

Oncogenes are a kind of abnormal genes that when they are activated, they predispose cells to develop into cancers[57,58,59,60]. They help to drive the rampant cell growth that underlies tumors. Fox example, when DNA damage occurs, the abnormal cells normally undergo a programmed form of death, know as apoptosis[61]. However, activated oncogenes may disturb the apoptosis procedure on these cells and trigger those cells to proliferate. Oncogenes are usually turned on by chromosomal translocations, copy number alterations or DNA damage, such as UV light, viruses or hazardous chemicals, and sometimes even by a missense point mutation in a gene[57,59]. Many oncogenes found in cancers are expressed at high levels or under focal amplifications[62]. Until now, hundreds of oncogenes have been identified in human cancer and the number continues to grow[63]. Some of these oncogene sequences or their products are considered as potential therapeutic targets for cancer drug development[64].

One of the well-documented oncogenes is *MYC*, which is frequently activated by missense mutations in many types of cancers[65,66]. It is located on the long arm of chromosome 8 near the telomere region. The discovery of *MYC* as a oncogene was first reported in a study of Burkitt's lymphoma in 1982[67,68,69]. In the following research, it was found to be

activated in many cancer types, including breast cancer, liver cancer, lung cancer, stomach cancer. *MYC* belongs to a family of transcription factors, which include *L-myc* [70,71] and *N-myc* [72,73] genes. The product of *MYC* is a hub protein that is at the center of a complex protein-protein interaction network, and participates in many growth promoting signal transduction pathways[74,75]. *MYC* is directly or indirectly involved in multiple cellular functions, including cell proliferation, growth, differentiation and apoptosis. *MYC* is also known to affect chromosomal stability, regulates a number of components of the mitotic checkpoint, and triggers metastasis[76]. Several mechanisms can regulate the level of expression of *MYC* by controlling its proximal promoter region[77].

On the other hand, tumor suppressor genes generally inhibit abnormal cell proliferation, which work the opposite way to oncogenes[78]. Loss or dysfunction of these genes will contribute to the development of cancer. Particularly, this kind of genes are usually involved in cell cycle checkpoint control, apoptosis promotion and DNA repair processes[79]. They help to prevent the accumulation of mutations in normal cells. In most cases, one copy of a tumor suppressor is sufficient to control cell proliferation. Thus, both alleles of a tumor suppressor gene must be lost or inactivated to contribute cancer development. This model is known as the two-hit hypothesis, and was first proposed in 1971 from the genetic mechanism analysis of retinoblastoma[80]. Today, this hypothesis serves as the basis for understanding the role of tumor suppressor genes in cancer development[81,82]. The inactivations may arise from point mutations, deletions or epigenetic silencing. Note that, there is an event, called “loss of heterozygosity” (LOH), that often leads to the inactivation of tumor suppressor gene[83,84]. During this process, in a heterozygous cell, one copy of the tumor suppressor gene is inactivated by loss-of-function mutations, which are usually point mutations or focal deletions. Subsequently, chromosomal deletions or somatic recombinations occur and replace the normal gene copy with a mutant copy. Furthermore, inherited tumor suppressor gene mutations are extremely useful in diagnostic applications. Tumor suppressor gene therapy was developed to treat cancer by restoring the function of a tumor suppressor gene lost or functionally inactivated in cancer cells[85,86]. The increasing knowledge of tumor suppressor genes will continue to enhance our ability to more successfully treat tumors at the molecular level.

The most representative tumor suppressor gene probably is *TP53*[87,88]. This gene is located on the short arm of chromosome 17. The involvement of *TP53* in neoplasia is more common than any other known tumor suppressors or oncogenes[89,90]. Therefore,

TP53 may have become the most striking tumor suppressor gene, which is believed to mutate in up to 50% of all human cancers[88]. The protein encoded by *TP53* is a highly linked node in the human protein-protein interaction network. *TP53* is involved in many cellular pathways, including DNA repair procedure, promote metastasis, trigger apoptosis and differentiation[91,92,93]. Loss of *p53* function is responsible for a number of human cancers. The dysfunction procedure can be caused by mutations of *p53* or perturbations in *p53* signaling pathways. Furthermore, the inherited germline mutations of a *TP53* allele will cause Li-Fraumeni syndrome (LFS)[94,95]. It is a disorder that significantly increases the risk of several types of cancer, e.g. leukemias, sarcomas, breast cancers and brain tumors[96]. In a number of clinical studies, a poor outcome has consistently been associated with mutated *TP53*. The growing understanding of the novel functions of *TP53* has already led to the identification of potential targets for new drugs effectively against cancer[97]. Table 4 is a list of important databases contain information about known cancer genes.

The correlation between cancer genes and CNAs is well defined, as discussed before. However, the relationship of CNAs to the gene density in the human genome is not determined yet. Our efforts of genomic array data collection enable us to perform a systematic study about this issue. The results of this study are presented in Chapter 5.

Table 4. Public databases that provide information on genes implicated in human cancer

Database	Description	Website
CancerGenes	A gene selection resource for cancer genome projects	cbio.mskcc.org
CancerResource	Cancer-relevant proteins and compound interactions	bioinf-data.charite.de/cancerresource/
CCDB	Cervical cancer gene database	crdd.osdd.net/raghava/ccdb/
CGED	Cancer gene expression database	lifesciencedb.jp/cged/
COLT-Cancer	Essential gene profiles in human cancer cell lines	colt.ccbr.utoronto.ca/cancer/
COSMIC	Catalogue Of Somatic Mutations In Cancer	cancer.sanger.ac.uk/cancergenome/projects/cosmic/
dbDEPC	Differentially Expressed Proteins in Human Cancer	lifecenter.sgst.cn/dbdepc
DDOC	Functional context of genes implicated in ovarian cancer	apps.sanbi.ac.za/ddoc/
DDPC	Database of Genes Associated with Prostate Cancer	apps.sanbi.ac.za/ddpc/
Network of Cancer Genes	Network properties of cancer genes	bio.ieo.eu/ncg/

2.3 EVOLUTION OF CANCER

Cancer is an evolutionary process[3]. The development of each tumor in different patients is a unique series of mutational events. It demonstrates the heterogeneity of cancer genome evolution. The occurrence of somatic mutations and natural selection are the underlying power of the evolutionary process. In the last decade, the application of high resolution microarray and massively parallel sequencing technologies provided new insights into the mutational landscape of cancer[98,99,100]. Through sequencing, the genomic sequence of tumor genome was compared to the sequence of corresponding germline DNA. This method generates comprehensive catalogues of somatic mutations in a single individual cancer genome.

Based on these valuable data, the classic model of carcinogenesis was proposed and comprehensively accepted[3,101]. In this model, random and heritable genomic mutations are accumulated in individual cells, which undergo natural selection. In the tumor microenvironment, cells with deleterious mutations are weeded out, whereas cells that have acquired specific alterations that promote proliferation will be preserved[3]. This competition process selects cells that survive more effectively than others. Over time and after many cell cycles, if a cell acquires a series of advantageous mutations that allow it to grow uncontrollably, to invade normal tissues, and to enable metastasis, it may gradually progress to invasive tumor. In this way, cancer is a multi-step procedure, and is the outcome of the Darwinian evolution process, like the events occur in species[3,102]. Getting a clear understanding of all the fundamental mutation events and their order during tumor evolution is important to elaborate the underlying mechanisms of cancer.

Since cancer evolution is a step-by-step process, it is believed that single mutations are not sufficient to cause cancer, but these alterations may initiate malignant growth or promote the cell's malignant transformation[2,3]. Although a large number of somatic abnormalities may occur in a cell's life cycle, it is not necessary that all these changes contribute to development of the cancer. Many, even most alterations have made no contribution at all. Given this situation, the concepts of "driver" and "passenger" mutations have been proposed[50,103].

Driver mutation is a kind of somatic mutations that disturb cancer genes and confer growth advantage on the cancer cell. They are positively selected in the environment of cancer

tissue during cancer evolution. Generally, driver mutations are required for maintenance of the tumor. In contrast with driver mutations, passenger mutations make no contribution to clonal growth and are not selected during cancer evolution[3,103]. These mutations are carried along with the driver mutation events and clonal expansion, and may be observed in the final tumor cells.

Obviously, it is important, also a challenge, to distinguish driver from passenger mutations in cancer research[50,104]. Although passenger mutations are likely to be randomly distributed in the genome, recurrent somatic mutation hotspots usually contain hundreds of genes. It again exhibits the heterogeneous nature of cancer genome.

Although the classic model of cancer progression has been comprehensively accepted, a phenomenon termed “chromothripsis” was recently characterized in cancer genomes, defined by the occurrence of tens to hundreds of clustered genomic rearrangements, supposedly having arisen in a single catastrophic event[105]. In this model, contiguous chromosomal regions are fragmented into many pieces, via presently unknown mechanisms[106,107]. Supposedly, these segments are then randomly fused together by the cell’s DNA repair machinery. Obviously, these somatically acquired genomic rearrangements may result in complex patterns of regional copy number changes. They have the potential to interrupt or activate multiple genes, and are consequently implicated in cancer development. It has been proposed that this “shattering” and aberrant repair of a multitude of DNA fragments may provide an alternative oncogenetic route, in contrast to the step-by-step paradigm of cancer development[105]. Figure 2 is a schema of chromothripsis that occurs in a single catastrophic chromosome shattering event. The initial study reported some evidence of a high prevalence in bone tumors[105]. The following studies of chromothripsis revealed that these events are also observed in multiple myeloma[108], colorectal cancer[109,110], acute lymphoblastic leukaemia[111], hepatocellular carcinoma[112], neuroblastoma[113], breast cancer[114,115] and medulloblastoma[116,117]. Besides human cancers, recent research have also reported chromothripsis events in germline[118,119] and non-human genomes[120].

The discovery of chromothripsis reveals a possible new paradigm in cancer development and stimulates great interest in the study of this phenomenon[121-132]. However, although several hypotheses have been proposed[105,107,125,133-136], the underlying mechanisms are still unknown. We launched a project that aims to integrate as much data as possible to

provide the most comprehensive overview of chromothripsis events, and eventually elucidating the potential mechanisms underlying chromothripsis. Since the chromosome-shattering was observed in a wide variety of tumors, the underlying mechanism is likely to reflect unknown general features of human cancer. The methods and results of this project are presented in Chapter 6.

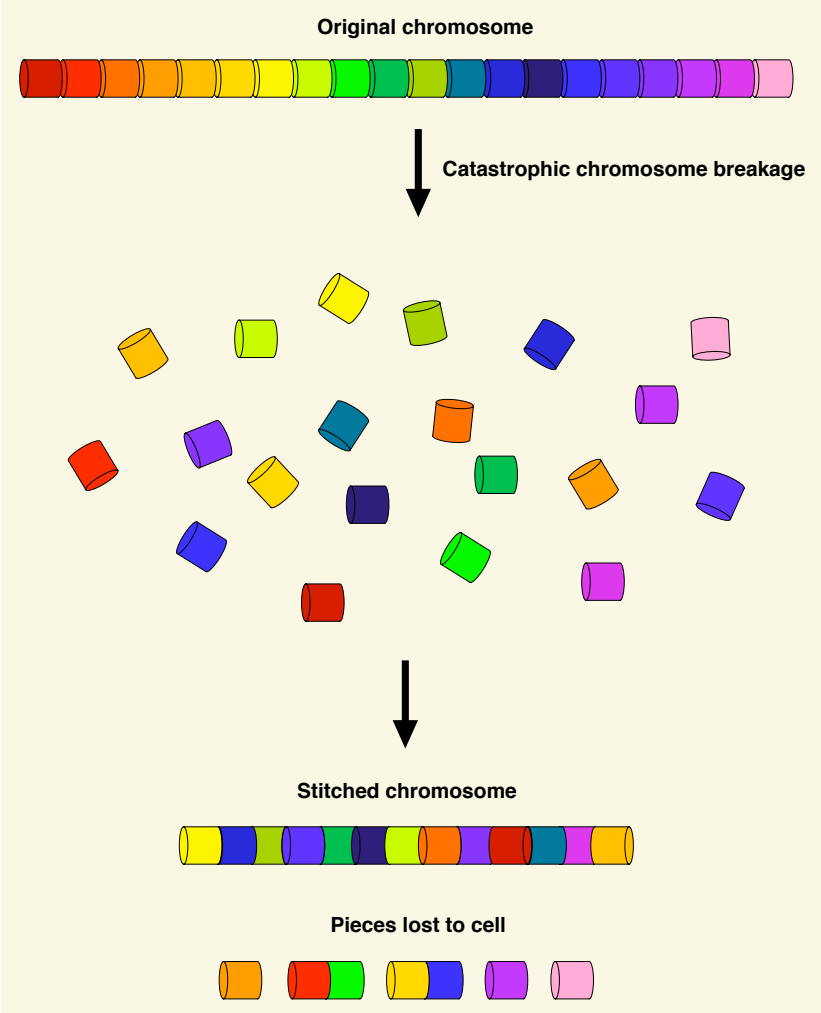


Figure 2. Schema of the process of chromothripsis in a single chromosome. The chromosome is shattered into many pieces. Then these fragments, or some of the fragments, are pieced together to generate a derivative chromosome. The other segments lost to cell. Chromothripsis can involve several chromosomes.

3 GENOMIC ARRAYS

3.1 EVOLUTION OF CYTOGENETICS AND MOLECULAR METHODS

The observed correlation between cancer initiation and/or progression and chromosome structural aberration has led to the development of several screening techniques capable of detecting DNA copy number changes. The conventional methods employed in cancer cytogenetics including chromosome banding and karyotyping[137,138,139]. These early techniques uncovered many genetic disorders caused by genome structural variation. The last two decade have witnessed substantial progress towards precise and genome-scale detection of copy number aberrations by molecular cytogenetic techniques[116,140,141]. By taking advantage of the unprecedented high resolution of these recently developed methods, nowadays we are able to extensively characterize chromosome rearrangements in tumor samples and reveal their functions in cancer development. The techniques applied in clinical diagnosis and research include fluorescence in situ hybridization[142,143,144] (FISH), chromosome- and array-based comparative genomic hybridization[11,55,145] (cCGH and aCGH), single-nucleotide polymorphisms[146,147] (SNP) arrays and next-generation sequencing[148,149] (NGS) methods. They completely revolutionized cytogenetic analyses and greatly stimulated interest in molecular cancer research. Figure 3 shows the evolution and the timeline of cytogenetic approaches.

In 1960, the Philadelphia (Ph) chromosome was identified as a prognostic marker among chronic myeloid leukemia (CML) patients[150]. It was the first time to successfully correlate a specific chromosomal abnormality with a particular malignant disease. Starting from this point, cytogenetics has become a powerful and promising tool in cancer research, especially after the invention of chromosome banding in 1970s[151,152]. In this technique, the metaphase chromosomes are stained with special dyes and then examined through a light or fluorescence microscope. In the metaphase stage of the cell cycle, chromatin is highly condensed and displays characteristic patterns of dark and bright bands. According to different staining methods, the resulting pattern can be classified according to the Q-, G- and R-banding nomenclature[153-156]. For example, the most commonly used Q- and G-

banding methods stain the chromosomes with quinacrine mustard and Giemsa staining, respectively. The dark G bands and bright Q bands are AT rich or gene-poor regions[157]. To the contrary, the G light and Q dark bands are GC rich or gene-rich sequences[157]. Based on the resolution of the technique, morphological regions can be divided into sub-bands. The regular banding produces about 550 light and dark bands, while the most high resolution banding produces 800 or more bands. With these bands, it is easy to ascertain chromosomal abnormalities in single cell, both numerical and structural, unbalanced and balanced. However, through limits in the spatial resolution when using these highly condensed chromosomes, metaphase banding is not sensitive enough to determine subtle aberrations (less than ~5 Mb). Also, due to the unspecific DNA staining chromosome banding is ineffective in characterizing complex rearrangements (those with more than 2 breakpoints). Even if it has these limitations, chromosome banding is very informative, and is still commonly used in diagnosis and as a prognostic marker[158,159].

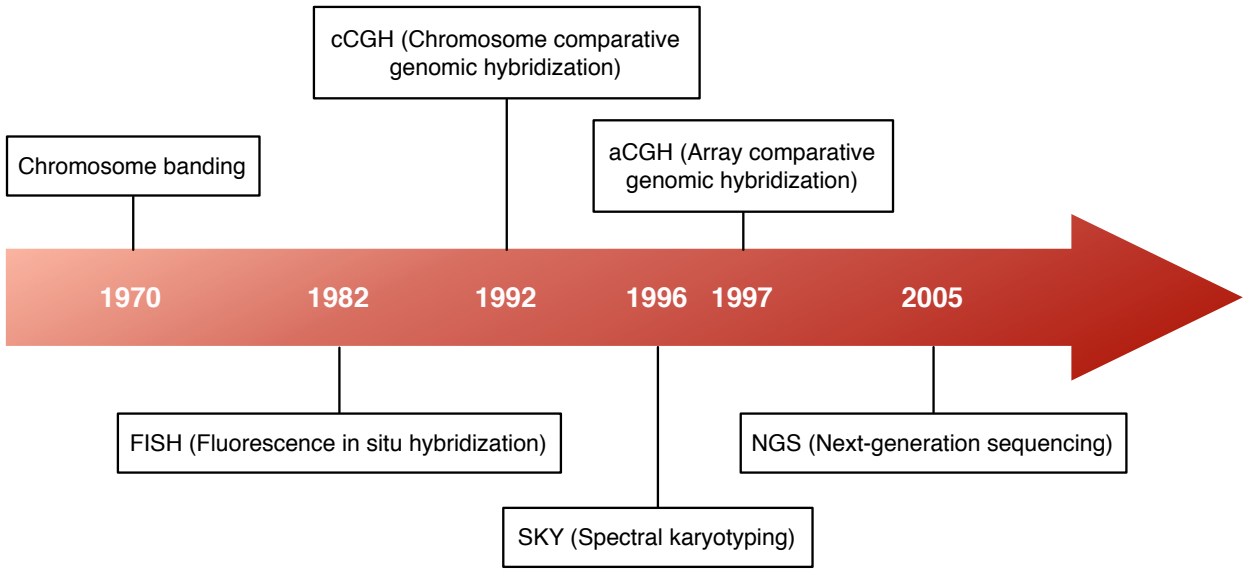


Figure 3. The evolution and timeline of cytogenetic methodologies. The "Chromosome banding" is the conventional cytogenetic method. The remaining approaches are molecular cytogenetic techniques, i.e. involve segment specific DNA labeling techniques.

Fluorescence in situ hybridization (FISH), introduced in the early 1980s, overcame the limitations of chromosome banding and became a significant diagnostic tool[142,144]. In brief, chromosome region specific DNA probes are labeled by fluorophores that can be directly detected by fluorescence microscopy. These probes are subsequently hybridized to their complementary DNA sequences in interphase cells or metaphase chromosomes. Fluorescent signal of interrogated specific genomic segments can be detected against a chromosomal background staining. The locations and number of these segments may reveal numerical or structural aberrations. Moreover, several DNA sequence targets can be concurrently detected in the same hybridization experiment by using probes labeled with different fluorescent dyes. FISH probes can even be manufactured with gene-level coverage, by using big plasmids such as cosmids/fosmids, or large bacterial artificial chromosomes[160,161,162] (BAC) or yeast artificial chromosome[163] (YAC). Furthermore, FISH is highly useful in mapping chromosome breakpoints, since chromosomes investigated at the metaphase or interphase stage of cell division may display split signals or aberrant fusion of differently labeled probes. As FISH analysis is considerably less affected by tissue processing artifacts, the results are widely accepted as the most reliable data and therefore generally regarded as “gold standard” for detecting specific chromosomal aberrations. However, a major disadvantage to perform FISH analysis is that prior knowledge of sequences to be examined is required. Although multicolor karyotyping can be applied to simultaneously analyzing several chromosomal loci at one time, the number of colors available for FISH is quite limited, and consequently it cannot be used for genome-wide screening of copy number alterations. Additionally, the probe preparation, labeling and signal evaluation are labor intensive and time consuming.

In 1992, a technique called comparative genomic hybridization (CGH) was introduced as a molecular tool in tumor cytogenetics[11]. It answers many deficiencies of conventional cytogenetics and FISH. This genome-wide technique allows the scanning of the whole genome for DNA deletions, duplications and amplifications in a single experiment. In this approach, genomic DNA is extracted from the cancer tissue sample and a normal reference sample, and differentially labeled with individual dyes, typically Cyanine 5 (red) and Cyanine 3 (green), respectively. The DNAs are mixed (1:1) and hybridized to normal human metaphase preparations on a microscope slide[11]. The red and green labelled DNA fragments compete for their complementary target sequences along each chromosome. The images of both fluorescent signals are captured, and the ratio of test to reference fluorescence along the chromosomes is quantified using digital image analysis. The red to

green fluorescence intensity ratios measured represent alterations of genetic material in the tumor at that specific locus[11]. For example, due to their equal coverage with hybridized DNA fragments from both samples, normal chromosomal regions would have a normalized ratio of one and a yellow coloration (pseudo-color) would be observed. Areas with deletions in the tumor genome would have fewer (or no) tumor DNA fragments competing for the respective template region and result in a ratio below one and an increased green signal from the reference DNA; amplified regions have a ratio above one, and would be revealed by relative increases in the red signal from the patient DNA. In this way, all over- and under-represented regions of each chromosome can be identified, and an overview of copy number changes throughout the whole tumor genome is obtained. The important advantages of this technique have enabled tremendous progress in the field of cancer research. First, chromosome-based CGH analysis can be performed without metaphase chromosomes from the tumor sample and therefore no cell culture is required. Second, it is a fast screening method and only relatively small amounts of DNA from the tumors are needed. Third, it enables analysis of the whole genome in a single experiment, and is more efficient in detecting the copy-number related result of complex structural aberrations.

Despite all these advantages, chromosome CGH does have some limitations that need to be taken into account. A main disadvantage is its inability to detect balanced abnormalities, such as translocations, inversions and polyploid changes. Moreover, the repetitive sequences in the probe have to be suppressed with Cot1 DNA (unlabeled repetitive DNA)[11,164] to reduce non-specific hybridization, or to be omitted in the subsequent interpretation process. Finally, the resolution of chromosome CGH is limited to 5-15 Mb, since the highly condensed metaphase chromosomes are used as the readout. Therefore, CNAs smaller than 10 Mb can not reliably be detected, which may limit the widespread clinical application of chromosome CGH. For the detection of such small CNAs, a high resolution technique, microarray-based CGH was developed with first applications reported in 1997 [55].

Like chromosome CGH, array CGH is a technique for genome-wide screening of copy number change events[55,145]. The difference is that the metaphase chromosomes are substituted by cloned DNA fragments spotted and immobilized onto glass microscope slides, or alternatively by oligonucleotide DNA sequences grown in situ on the array substrate. These fragments are microarray probes of which the exact genomic positions

are known, and are designed in various sizes for the different platforms. Similar to the chromosome CGH procedure, equal amounts of patient and reference DNAs are extracted and differentially labeled with distinctive fluorochromes[55,145]. DNAs from both sources are co-hybridized to a glass slide containing the array probes. A microarray scanner is used to scan the slide and convert image files into fluorescent intensity signals. For each single clone, the test to control fluorescence signal ratio is proportional to the ratio of the copy numbers of related location in the patient and reference genomes. The output of thousands of spots with different ratios of the signals provide a high resolution overview of CNAs throughout the whole genome. The schematic overview of the array CGH technique is shown in Figure 4. The reference sample can be the normal tissue of the same patient or pooled genomic DNA. In array CGH, the genomic resolution of different platforms is determined by the distance and size of the DNA probes. These interrogating probes can be prepared by bacterial artificial chromosomes (BAC) or P1 (PAC) clones[165], cosmids and fosmids, or based on single-stranded oligonucleotides[166]. Regardless of the many types of probes, the significant advantage of array CGH over chromosome CGH is the high resolution and accordingly, the ability to precisely map copy number aberrations. It has been proven that the use of array CGH has revealed a number of new recurring genomic abnormalities that can not be identified by chromosome CGH. The highly improved resolution achieved by array CGH enables this technique to detect the precise boundaries of CNAs, which may disturb oncogenes or tumor suppressors, and provide insights into the mechanisms underlying CNA formation. Moreover, array CGH is sensitive enough for studying complex rearrangements generated by more than two chromosome breakpoints[167]. Thus, it quickly replaced chromosome CGH in the last decade, becoming ubiquitous tools in cancer research as well as human genetic disorder studies[168].

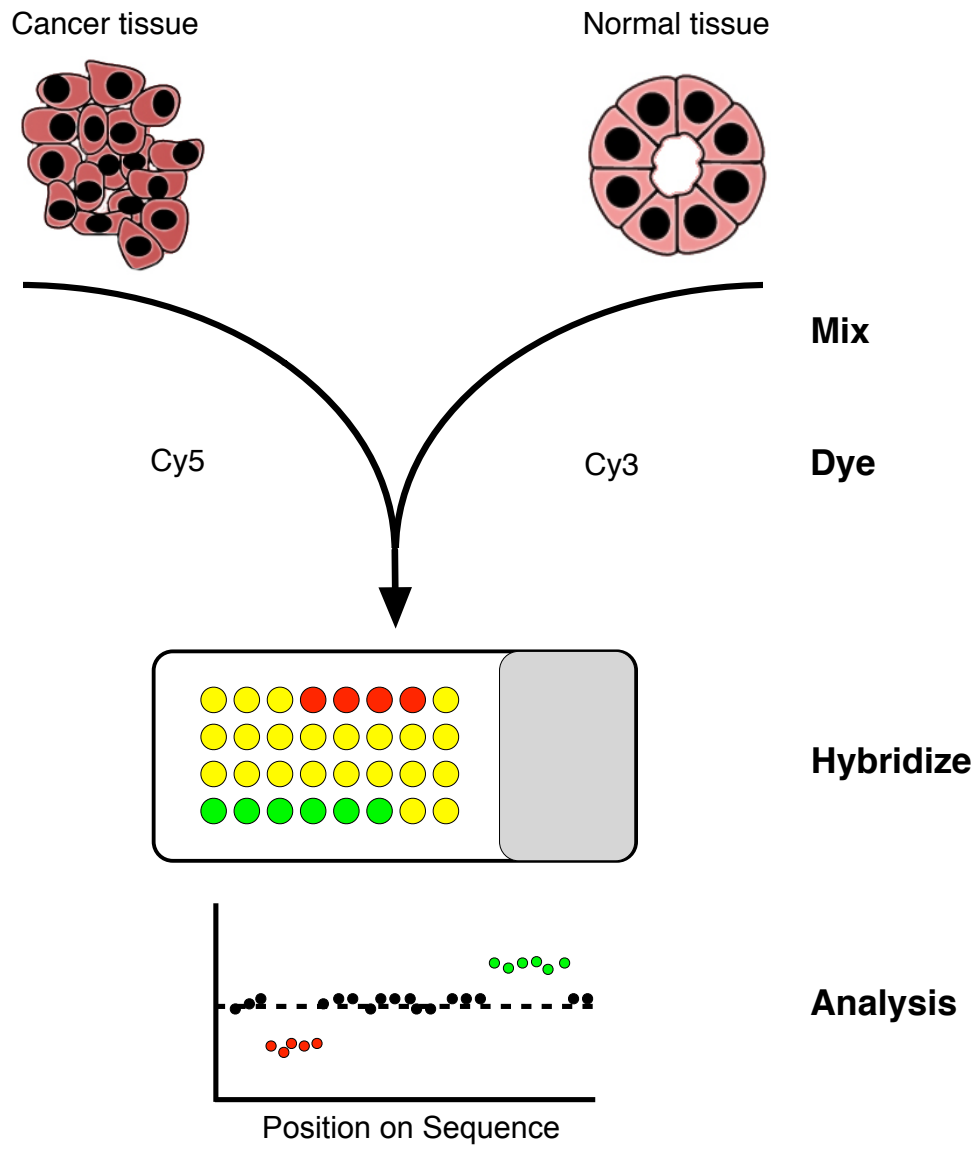


Figure 4. Overview of array CGH technology. Genomic DNAs from cancer samples and references are differentially labeled with green (Cy3) and red (Cy5) fluorescence dyes, respectively (Other fluorochromes can also be used, like fluorescein isothiocyanate). The two samples are mixed and competitively co-hybridized to an array containing DNA sequence targets that have been spotted on a glass slide. The schema of the slide depicts hundreds of spots with different ratios of the fluorescence intensities. These values are proportional to the ratio of the copy numbers of genomic DNA in the tumor samples and reference genomes. The yellow spots on the slide indicate the presence of equal amounts of tumor and normal DNA. The spots that appear green indicate extra chromosomal material in the tumor sample at that specific location. The red spots indicate loss of tumor DNA in the sample. The Cy3/Cy5 fluorescent ratio on each spot is captured and normalized to Log2 values, and the median ratio is 0. In the ideogram of signal intensity plot, each dot represents a single probe that spotted on the array, and the ratio for each probe is plotted in genome order. The five red probes have a ratio near -1, indicating a genomic loss. The six green probes have a ratio near 1, indicating a genomic gain. These results can be converted to a high resolution overview of copy number changes throughout the whole genome.

As with other cytogenetic methodologies, array CGH technology also has a number of limitations. It is not able to identify balanced rearrangements. Like in chromosome CGH, only copy number imbalances can be detected. Furthermore, it reflects an average pattern of all sub-populations in a tumor sample, so that intercellular variabilities are not easily assessed.

Figure 5 shows the publication statistics for studies performed chromosome-based and array-based CGH[169]. In the last decade, there has been a significant gradual shift of interest from cCGH to aCGH. Array CGH data analysis and interpretation is the major focus of this thesis. Since there are so many different types of platforms in aCGH technique, and each type has its unique advantages and disadvantages, in the following, different platforms will be introduced in detail.

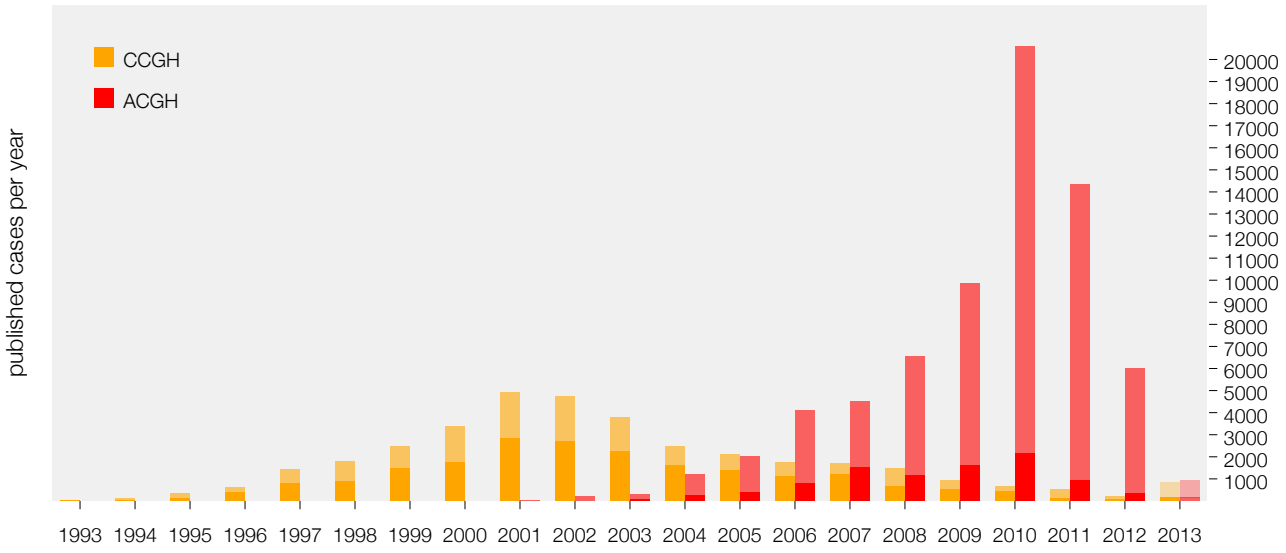


Figure 5. The statistics for published chromosome CGH data and oncogenomic arrays. The y-axis represents the number of individual tumors reported in the literature. The data shows a growing interest in applying array CGH techniques for feature detection in cancer samples. The shaded bars correspond to the total number of cases discussed in the articles, and the solid bars show raw or annotated data that can be collected. Data is obtained from Progenetix database and up to February 2013.

3.2 LARGE CLONE ARRAYS

The first array CGH platform was introduced in 1997 with the purpose of enhancing the resolution and simplifying the analysis procedure of chromosome CGH[55]. In these arrays, in place of normal metaphase chromosomes thousands of target DNA sequences, i.e. array probes, were immobilized on glass slides. These sequences were largely cloned in bacterial artificial chromosome or P1-derived artificial chromosome vectors, of typically 75-130 kb in length. Other clone types used include cosmids and fosmids.

The resolution of platforms is mainly determined by the length of probe and the density of genome coverage by the clones on the array. Initially, BAC arrays are comprised of ~3000 large insert clones spaced at ~1 Mb intervals throughout the genome[170]. They can reliably detect larger copy number aberrations, but genomic changes that occur in the gaps between the probes will not be identified. Tiling path BAC arrays comprised of 32 000 or more clones covering the entire genome were designed later and have allowed for a dramatic increase in resolution[171]. In this way, the theoretical resolution is only limited by the size of the probes used.

Besides their resolution, another important parameter of array platforms, is their sensitivity. It is partly influenced by the inherent biological variability of the samples used in each study. For example, due to the lack of non-cancer cells (e.g. stroma cells, blood vessels etc.), cell lines offer a more homogeneous cell source than primary samples. Primary tissues are frequently contaminated by normal cells, which will reduce the hybridization ratio of CNAs. Furthermore, sub-clones of cells that arise during tumor evolution will affect the magnitude of specific changes in the profile generated by arrays. Therefore, the primary tumors give the overall view of multiple heterogeneous subclones, and the contribution of minor variant tumor cells may not reliably be observed.

Among all the oncogenomic platforms, large clone arrays provide the best signal to noise ratios. They allow highly sensitive and reproducible detection of a wide range of copy number changes including single copy number gains/losses and high-level amplifications[164]. Moreover, since their probes are long enough to minimize non-specific hybridization across the array, a CNA that involves only a single probe can be considered as a real genetic event in the tumor instead of an artifact of the technique, if concerning a well-defined target. However, one major drawback of large clone arrays, as well as most

array CGH platforms, is the inability to detect rearrangements without change in DNA content, such as loss of heterozygosity (LOH) and translocations. Thus, they will have limited use in diagnosis of, for instance, hematologic malignancies, for which copy-neutral aberrations (e.g. fusion gene generation) known as one of the important mechanisms.

For large clone arrays, there are not a wide variety of commercial alternatives available. The most commonly used are in-house printed arrays, typically UCSF HumArray, VU University Medical Center Microarray Core Facility BAC array, DKFZ (The German Cancer Research Center) Human BAC/P1 array, etc. These arrays are generally cheaper than commercial arrays.

3.3 CDNA ARRAYS

Soon after the development of BAC/P1 arrays, ~30 000 accurately mapped cDNAs were also used as genomic resource for mapping DNA copy number aberrations across a whole genome[172,173]. Although cDNA arrays were originally designed for analysis of gene expression levels, they in principle provided a further increase in resolution of aCGH compared to then available BAC/P1 platforms.

In this technique, DNA copy number changes can be measured in a gene specific manner. Disease-related genes, such as oncogenes and tumor suppressor genes, can be selectively spotted on the array to provide important information. Moreover, the potential advantage of using cDNA arrays is that expression analysis of a given sample can be carried out in parallel with DNA copy number characterization, on the same platform. The combination of copy number and gene expression patterns should prove useful for identifying pathogenically important genes in amplicons, in that it can confirm the copy number induced expression change in the genes under consideration. In cancer research, this feature may facilitate to distinguish cancer “driver” genes from co-amplified “passenger” genes. Following the completion and annotation of human genome sequence, this platform was thought to greatly accelerate the discovery of genes that play roles in tumor progression.

However, cDNA arrays have unique disadvantages compared to other platforms[170,172,173]. First, they are not possible to pre-select sequences of cDNAs, and therefore the opportunity for discovery novel aberrations is limited. Second, since genes are not uniformly distributed throughout the genome, it is difficult to define the resolution of cDNA arrays. The resolution can be variable across the gene-rich and gene-poor regions. This may increase the uncertainty in the identification of the exact location of breakpoints. Third, cDNAs may contain repetitive sequences or short sequences which share extensive homology. This phenomenon may influence the signal intensity of specific probe and lead to mis-detection of abnormal regions.

The cDNA arrays were first introduced by Stanford University, and oligonucleotides are printed on glass slides using a robotic applicator with multiple capillary needles[173]. These arrays have been widely used in clinical applications because of their relative affordability in terms of cost. Beyond this, no specific equipment is required for carrying out hybridization and data capture. The most common platforms are SMD 40k cDNA arrays.

3.4 OLIGONUCLEOTIDE ARRAYS

Oligonucleotide-based arrays hold the potential of enhanced design flexibility and to overcome some of the drawbacks of the aforementioned techniques[174,175,176]. Unlike cDNA platforms, oligonucleotide arrays are able to distinguish between highly homologous sequences by careful design, as the sequence of probes can be predetermined. A high dense coverage of the genome and relatively high resolution can be obtained by the large number of probes.

Oligonucleotide probes can be roughly classified into the “short” and “long” categories. Short oligonucleotide arrays (21-25 mers) were originally designed to characterize single nucleotide polymorphisms (SNPs) in the genome as represented e.g. by the Affymetrix platforms[177]. These arrays have been recruited for assessing copy number aberrations in CGH experiments, and will be described in detail in the following section. The long oligonucleotide probes are typically 60-70 mers, resulting in a higher probability of non-repetitive sequences and improved hybridization specificity[174]. The fluorescence intensities

detected on a genome-wide scale are comparable to BAC/P1 arrays in the magnitude of signal and background noise.

As a special example, ROMA (representational oligonucleotide microarray analysis)[178] is a kind of platform that reduces sample genome complexity and hybridize to long oligonucleotide probe sets. Before hybridization, DNA samples have to be cleaved with restriction enzyme, and followed by linker-mediated PCR that converts the genomic DNA into a predictable sample of reduced complexity. In this way, about 2.5% genome is covered, which improves the signal to noise ratio. After amplifications and with sufficient probe density, single copy number aberrations can be observed. However, although ROMA allows a genome-wide identification of CNAs, the low coverage of the genome may lead to large gaps where no genomic information is available. The whole-genome tiling path arrays should avoid this problem and provide a higher resolution.

Today a great variety of commercial oligonucleotide platforms are available. The Agilent arrays carry 60 mers probes and have been successfully used for clinical cytogenetic diagnosis[177,179,180]. The most popular platforms here are 4x44K, 4x180K and 244K. NimbleGen arrays carry 45-85 mers probes and provide flexible design. NimbleGen 2.1M arrays contain more probes that allow the researchers to examine focal CNAs at high resolution, and other platforms including 85K, 385K, 720K CGH arrays etc. Illumina has developed a novel bead array technology which is utilized for a broad range of DNA and RNA analysis applications. The common CGH platforms are BeadChip 317K, 610Quad and 660Quad. The Affymetrix arrays use photolithography to build 25 mers oligonucleotides directly on the array slides. Their outstanding sensitivity and accuracy is achieved by the high density of short oligo probes, and will be discussed in the following part.

3.5 SNP ARRAYS

Single nucleotide polymorphism arrays are one type of oligonucleotide arrays, which were originally designed to detect common SNPs and were mainly used in genotyping studies[181,182,183]. In addition, these platforms are also used to perform copy number

abnormality identification. Unlike the above-mentioned arrays, which rely on co-hybridization of patient and reference DNA, only a single hybridization is performed for patient samples[184,185]. The observed tumor sample intensities are compared to expected reference intensities, which can be generated by the same laboratory or from other archive sources such as the HapMap Project[186,187] (The International HapMap Consortium). However, the reference dataset used may influence the quality control procedure or the sensitivity of detecting CNAs. According to previous studies, a large in-house reference dataset is more likely to produce improved quality of results in comparison with the external reference datasets. With the normalized Log2 intensity ratios, genomic gains and losses can then be called. The relative short probes, usually 25 mers oligonucleotides, used for SNP detection provide the lower noise-to-signal ratios than longer probes. Therefore, several probes are required to define a copy number alteration event. In general, the design of the original SNP arrays focused on specific loci in order to detect SNPs, instead of providing an even coverage of the genome.

Besides genomic gains and losses that can be detected by both CGH and SNP arrays, SNP platforms offer the unique possibility to simultaneously analyze loss of heterozygosity using the same array[183]. The signals from individual alleles can be detected to calculate an allelic ratio (AR). The AR patterns may reveal particular classes of CNAs. Thus, it is possible to distinguish copy number loss from copy neutral genetic events underlying LOH. This may be of significance for cancer research, since LOH regions may also correlate with increased expression of oncogenes or inactivation of tumor suppressors. For instance, the characterization of haplotype structures could facilitate the analysis of cancer predisposition.

One important capability of SNP arrays is to discovery CNVs. To improve the efficiency of CNV detection, non-genotyping probes are also included in recent SNP arrays versions, and the overall probe density is also increasing[183,188,189]. For example, in the Affymetrix SNP6 platform which includes 1.8 million probes, both polymorphic or nonpolymorphic probes are combined to detect CNVs, copy number aberrations and SNPs.

Another potential function of SNP platforms is to detect polyploidy. This phenomenon is observed in most cancer types and supposed to be an important mechanism in tumor development[183,189]. In some cases, the polyploid tumor samples are not possible to be identified by the overall hybridization signal on arrays due to experimental and

computational normalization to a virtual diploid state, and only relative copy number changes against the baseline are able to be detected. The integration of allele specific signals and Log2 intensity ratios allows us to assign absolute copy numbers to investigated genomic regions and determine if samples are polyploidy[190,191].

With the unique advantages of being able to detect copy neutral LOH, uniparental disomy and regions identical by descent, SNP arrays have undergone huge developments over the last decade. So far, commercial SNP arrays have achieved remarkable success. Affymetrix Microarrays provides robust precise platforms with various resolutions for detecting CNAs and SNPs throughout the whole genome. Affymetrix 100K Mapping arrays have a mean marker distance of 24kb, and 500K Mapping arrays have an average inter-probe distance of 5.8 kb. Recently, Agilent provided “CGH + SNP” arrays that are also designed both for capture of copy number changes and for detection of LOH. The available platforms including Agilent SurePrint G3 Human Genome CGH+SNP 2x400K and 4x180K Microarray. Currently available “SNP” platforms with the highest density can analyze more than 2 million loci using a single array.

Part II

RESULTS

4 CURATED DATABASE FOR COPY NUMBER PROFILING DATA IN HUMAN CANCER

4.1 PREFACE

Genomic copy number aberrations are a hallmark of almost all forms of human malignancies[3]. The recurrent genomic imbalances may reveal mutations in oncogenes or tumor suppressors[45]. CNA hot spots identification has proven to be efficient to reveal novel cancer-causing genes[98]. Besides genes, on a systems-level, CNA data is used to detect pathways altered in cancers and to deduce functional relevance of pathway members[98]. Since specific CNAs may be attributed to certain tumor types, in some studies copy number profiling is employed to differentiate biological as well as clinical subtypes. Subtype-associated CNA regions facilitate to understand biological differences and lead to discovery of new therapeutic targets. Moreover, gene expression or somatic mutation data offer complementary perspectives on the same cancer genome. Thus, CNAs can be integrated with these data to provide a more comprehensive model of cancer development[3].

In the last two decades, molecular-cytogenetic techniques have been applied to screen DNA copy number profiles in human cancer[98,131]. Among these techniques, chromosomal and array comparative genomic hybridization were comprehensively employed for genome-wide analysis. While chromosomal CGH has a limited spatial resolution of several megabases, the resolution of array based technologies is mainly limited due to cost/benefit evaluations instead of technical obstacles. The flood of high resolution aCGH data have led to an increased interest in genetic and cancer research, also offer new challenges and opportunities for large-scale genomic data mining, modeling and integration. As a result, a more complete understanding of functional effect of CNAs in the context of cancer can be obtained.

Several online oncogenomic resources have been launched, focusing on different aspects of data content as well as representation. In principle, these databases facilitate access

and visualization of CNA data. However, they are limited to specific aCGH platforms or single institutions as well as limited disease categories. Several commonly used genomic array databases are listed below.

• National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO)

GEO is the largest international public repository for raw high-throughput microarray data[26]. At the moment, there are more than 32000 series, comprising 800000 samples derived from ~1600 organisms included in GEO database. These data sets were submitted by the research community. Although GEO provides several web-based tools and strategies to assist users to query and analyze data, the main principle of this database is to archive raw data. It is not able to allow users to intuitively visualize array data.

• European Bioinformatics Institute (EBI) ArrayExpress

ArrayExpress is another major international repository for functional genomics data mainly from microarray[27]. The database supports data query and download. Data sets were either submitted directly by the research community or were imported from the NCBI GEO. It currently contains data from about one million entities and over 30000 experiments. Although ArryExpress provides several pre-analyzed data formats, such as R objects for array data, this database also aims to archive raw array data. Thus, it is not easy for users to perform custom data analysis and visualization.

• The Cancer Genome Atlas (TCGA)

TCGA dedicates efforts to understand the molecular basis of cancer through large-scale genomic data[28]. Immense amounts of array data were integrated into the database. In addition, it provides informatics tools to the research community to make use of these data. At the moment, about 7000 arrays from 26 cancer types are stored in TCGA. However, only segmented data is publicly available, while the raw data is protected.

• CaSNP

CaSNP database focuses on the collection of CNA information from single-nucleotide polymorphism arrays[192]. It contains ~11500 SNP array data in 104 series. All the data were obtained from GEO. CaSNP allows users to input region or gene of interest, and the

summarized frequencies of genomic gain and loss will be returned. It also displays the heatmap of copy numbers estimated at each SNP maker. CaSNP is a useful tool for cancer CNA research. However, it is limited to SNP arrays with single data source.

• **CanGem**

CanGem is a public web-based database for storing genomic array data[193]. Additionally, it collects clinical information whenever it is available. CanGem provides custom datasets download by querying for specific clinical sample information or copy number changes of individual genes. It provides more than 1000 arrays from different data sources.

Although the above-mentioned data sources provide large amount of array data, no single data source does yet provide an extensive collection of oncogenomic CNA data, which readily could be used for genomic feature mining, across a representative range of cancer entities. A comprehensive reference database for copy number profiling data in human tumor is needed to promote more effective cancer research.

4.2 arrayMap

For this research project, I collected array CGH data from 4 main data sources, including GEO, ArrayExpress, TCGA and publication supplements, and created arrayMap database for providing genomic copy number data sets in human cancer[23] (Figure 6). I set up a pipeline to accumulate and process oncogenomic array data into a unified and structured format. Associated histopathological and clinical information were incorporated into the database. So far, arrayMap contains more than 40,000 arrays on 197 cancer types. Figure 7 shows the summary of CNA data stored in arrayMap. Samples of interest can be browsed, visualized and analyzed via an intuitive interface. Computational tools are provided for biostatistical data analysis, such as CNA clustering for case specific or for subset data and basic clinical correlations. So far as I know, arrayMap provides the largest annotated data for whole genome CNA profiles. arrayMap is publicly available at www.arraymap.org.

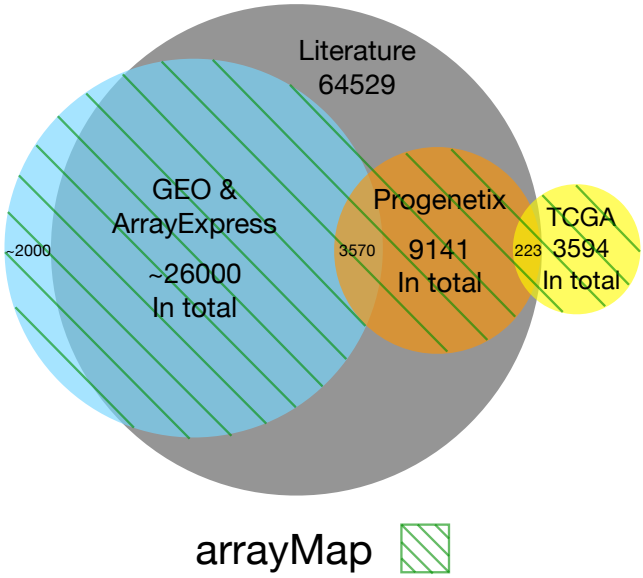


Figure 6. An overview of data sources included in arrayMap. The core data of arrayMap is obtained from GEO and ArrayExpress. According to our data collection efforts, there are totally about 65000 published aCGH cases in literatures. In arrayMap, there are many arrays that are submitted into GEO/ArrayExpress or TCGA, but are not published yet.

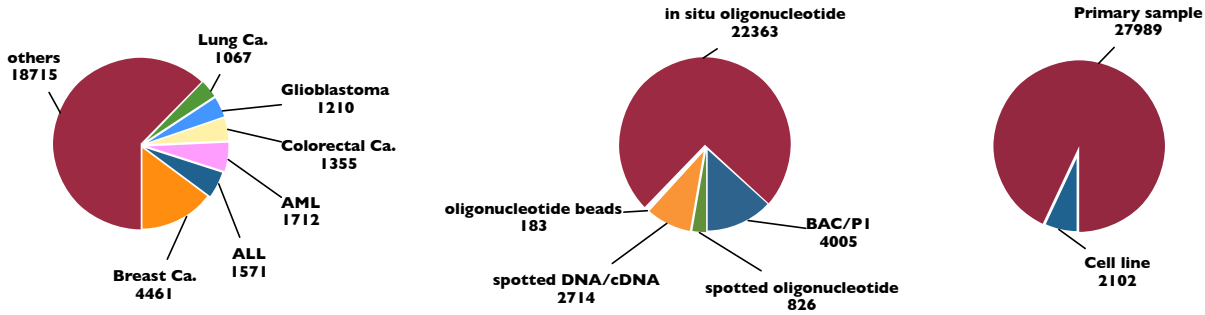


Figure 7. Summary of oncogenomic data stored in arrayMap. The numbers represent tumor samples. AML, acute myeloid leukemia; ALL, acute lymphoblastic leukemia.

The publication is included below.

OPEN ACCESS Freely available online

PLoS one

arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies

Haoyang Cai[•], **Nitin Kumar**[•], **Michael Baudis***

Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Abstract

Background: The delineation of genomic copy number abnormalities (CNAs) from cancer samples has been instrumental for identification of tumor suppressor genes and oncogenes and proven useful for clinical marker detection. An increasing number of projects have mapped CNAs using high-resolution microarray based techniques. So far, no single resource does provide a global collection of readily accessible oncogenomic array data.

Methodology/Principal Findings: We here present arrayMap, a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides a platform for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data. To date, the resource incorporates more than 40,000 arrays in 224 cancer types extracted from several resources, including the NCBI's Gene Expression Omnibus (GEO), EBI's ArrayExpress (AE), The Cancer Genome Atlas (TCGA), publication supplements and direct submissions. For the majority of the included datasets, probe level and integrated visualization facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools.

Conclusions/Significance: To our knowledge, currently no data source provides an extensive collection of high resolution oncogenomic CNA data which readily could be used for genomic feature mining, across a representative range of cancer entities. arrayMap represents our effort for providing a long term platform for oncogenomic CNA data independent of specific platform considerations or specific project dependence. The online database can be accessed at http//www.arraymap.org.

Citation: Cai H, Kumar N, Baudis M (2012) arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. PLoS ONE 7(5): e36944. doi:10.1371/journal.pone.0036944

Editor: Ying Xu, University of Georgia, United States of America

Received January 10, 2012; **Accepted** April 16, 2012; **Published** May 18, 2012

Funding: HC is supported through a personal grant from the China Scholarship Council. NK and MB had received support through the Krebsliga Schweiz and the University of Zurich's Research Priority Program Systems Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: michael.baudis@imls.uzh.ch

• These authors contributed equally to this work.

Introduction

Genomic copy number abnormalities (CNAs) are a relevant feature in the development of basically all forms of human malignancies [1]. Many genomic imbalances are recurrent and display tumor-specific patterns [2,3]. It is believed that these genomic instabilities reveal mutations in tumor suppressor genes and oncogenes which eventually result in a clone of fully malignant cells. Investigation of CNA hot spots (chromosomal loci frequently involved in CNA) has proven to be an effective methodology to identify novel cancer-causing genes [4,5]. On a systems level, CNA data along with expression or somatic mutation data is used to detect pathways altered in cancers and to deduce functional relevance of pathway members [6,7]. Since many CNAs have been attributed to specific tumor types or clinical risk profiles, in some entities copy number profiling is employed to characterize different biological as well as clinical subtypes with implications for treatment and individual prognosis. Subtype-associated CNA regions are used to predict causative genes, furthering understanding of biological differences and leading to discovery of new therapeutic targets [8,9].

Throughout the last two decades, molecular-cytogenetic techniques have been applied to scan genomic copy number profiles in virtually all types of human neoplasias. For whole genome analysis, these techniques predominantly consist of chromosomal and array comparative genomic hybridization (CGH), including CNA detection by cDNA and single nucleotide polymorphism (SNP) arrays [10–12]. While chromosomal CGH has a limited spatial resolution of several megabases, the resolution of recent array based technologies (aCGH) is mainly limited due to cost/benefit evaluations instead of technical obstacles. In this article, we use the terms "array CGH" and "aCGH" for all technical variants of whole genome copy number arrays. This includes e.g. single color arrays for which regional copy number normalization is performed through bioinformatics procedures applied to external references and internal data distribution.

The flood of new insights into structural genomic changes in health and disease has led to an increased interest in genomic data sets in genetic and cancer research. Several systematic studies of CNAs across many cancer types have been performed [13,14].

These efforts attempt a more complete understanding of functional effect of CNAs in the context of cancer.

The exponential increase of high resolution CNA datasets offers new challenges and opportunities for large-scale genomic data mining, data modeling and functional data integration. Several online resources have been developed, focusing on different aspects of data content as well as representation [6,15–19]. An overview of some of the prominent examples is given in Table 1. In principle, these databases facilitate access and utilization of CNA data. However, they are limited to specific aCGH platforms and/or single institutions as well as limited disease categories, or, as in the cases of GEO [15] and Ensembl ArrayExpress [16], mainly serve as raw data repositories. To the best of our knowledge, no single data source does yet provide an extensive collection of high resolution oncogenomic CNA data which readily could be used for genomic feature mining, across a representative range of cancer entities.

Here we present "arrayMap", a web-based reference database for genomic copy number data sets in cancer. We have generated a pipeline to accumulate and process oncogenomic array data into a unified and structured format. The resource incorporates associated histopathological and clinical information where accessible.

So far, arrayMap contains more than 40,000 arrays on 224 cancer types from five main data sources: NCBI GEO, EBI ArrayExpress, The Cancer Genome Atlas, publication supplements and user submitted data. Samples of interest can be browsed, visualized and analyzed via an intuitive interface. Computational tools are provided for biostatistical data analysis such as CNA clustering for case specific or for subset data and basic clinical correlations. arrayMap is publicly available at www.arraymap.org.

Results

Data Content

Our combination of both "top-down" (publication driven) as well as "bottom-up" (array data driven) approaches allowed us to identify a comprehensive set of accessible aCGH based cancer CNA data sets and to estimate the ratio of accessible data of the overall published/deposited data.

As main result of the array data driven approach, we extracted 495 series comprising of 32002 arrays, generated on 237 platforms from NCBI's GEO. Among those, raw data files of approximately 29000 whole genome arrays were suitable for inclusion into our data processing pipeline. When reviewing the content of AE, we found that the majority of AE cancer genome data sets were also submitted to GEO. At the time of writing, 11 datasets including 712 arrays not present in GEO had been processed based on AE specific series. Detailed information on the GEO/AE data sets is provided in Table S1.

The top-down procedure was based on our group's continuous monitoring of cancer related articles utilizing genome copy number screening approaches, as established for our "Progenetix" project (www.progenetix.org; [19]). The census date for the literature based data collection was August 15 2011. At this point, we had identified 931 articles discussing a total of 53213 genomic cancer CNA profiles based on aCGH techniques. Of these, 8728 cases out of 199 articles so far had been extracted from publication related sources (e.g. supplementary data tables) and annotated and made been accessible through Progenetix. This data included cases for which only supervised information but no probe data was available (e.g. author annotated Golden Path or cytogenetic CNA regions). Literature based data sets containing probe specific data or with the respective data presented to us by the authors (640 samples) were included into our arrayMap data processing pipeline.

The data content of arrayMap is summarized in Table 2. Current numbers on the website will include changes based on ongoing annotation efforts (i.e. addition of data sets, removal of low quality arrays).

As a by-product of our data collection and annotation efforts, we are able to provide estimates of content and trends for the platform usage and cancer entity coverage for the majority of published data. According to the assigned ICD-O 3 (International Classification of Diseases for Oncology, 3rd Edition) code and descriptive diagnostic text, breast carcinoma predominates as single largest clinical entity with 6459 arrays. Table S2 presents sample sets in arrayMap classified by ICD-O code.

The most widely available array CGH platforms are either based on large insert clones (BAC/P1 arrays) or based on shorter single-stranded DNA molecules (oligonucleotide arrays), which may or may not include single-nucleotide polymorphism specific probe sequences (SNP arrays). Also, although designed for gene expression profiling, cDNA arrays were used by several laboratories for measuring genomic copy number changes. Although all these platforms are considered suitable for whole genome CNA analysis, their probe densities and other parameters can affect specific features of the analysis results [20–23]. Table S3 lists the general platform types and corresponding overall numbers of the data registered in arrayMap.

Table 1. Prominent online resources of genomic data.

Name	Address	Platform(s)	Data format	Comment
GEO [15]	www.ncbi.nlm.nih.gov/geo	263	raw & normalized probe signal intensity	largest microarray data repository
ArrayExpress* [16]	www.ebi.ac.uk/arrayexpress	16	raw & normalized probe signal intensity	many duplicate data in GEO
TCGA [6]	cancergenome.nih.gov	1	segmentation data	raw probe data is limited to download
CanGEM** [17]	www.cangem.org	38	normalized probe signal intensity	including many types of microarray data
CaSNP [18]	cistrome.dfci.harvard.edu/CaSNP	8	average copy number & graphic	focus on SNP array data
Progenetix [19]	www.progenetix.org	235	ISCN*** & golden path	data from publications

Data up to 29 April, 2011.
*excluding data both in GEO and ArrayExpress.
**statistical information only including CGH, SNP and cDNA data.
***International system for human cytogenetic nomenclature.
doi:10.1371/journal.pone.0036944.t001

Table 2. aCGH data integrated in arrayMap.

Data Source	Arrays	Cases	Series	Platforms	Publications
GEO	32002	25728	495	237	490
ArrayExpress	712		11	16	11
TCGA	7249	3594	19	1	*
Publication Supplements	>4578**	4578			137
Author Submission	556	539	8	7	

Data up to 29 April, 2011.
*Due to lack of publication information, there may be a small amount of duplicate data in GEO.
**Array number may be higher than case number since reported results per case occasionally may be based on more than one array. The number does not include data presented both in publication supplements as well as GEO.
doi:10.1371/journal.pone.0036944.t002

In reviewing the technical platform composition, two related trends become apparent (Figure 1). Originally developed in groups with expertise in molecular cytogenetics and cancer genome analysis, printed large insert clone arrays (BAC/P1) were the first whole genome CNA screening tools with a spatial resolution surpassing that of chromosomal CGH. Other groups re-employed cDNA arrays, developed for expression screening, for genomic hybridizations. However, over the last years one can observe the overwhelming use of various industrially produced oligonucleotide array platforms, which compensate their low single probe fidelity through a probe density at 1–3 orders of magnitude higher than common for BAC/P1 arrays. Another reason for the success of oligonucleotide arrays is the integration of SNP specific probes, which in principle allows to use of the same experiments for genetic association studies and the evaluation of copy number neutral loss of heterozygosity regions [12,24,25].

Data Access and Usage Scenarios

Based on our experience from the Progenetix project, a strong emphasis was put on a user friendly data interface. Here, we followed a ''dual user type'' scenario: Users without bioinformatics background should be able to intuitively visualize core data features as well as to perform standard analysis procedures, while for bioinformaticians the formatted database content should be accessible to use with their analysis tools of choice.

Query interface. Data browsing in arrayMap is based on two types of query methods: search by experimental series metadata and search by sample features.

In the series query form, users can perform various search options by specifying (i) descriptive diagnosis text; (ii) disease classification (ICD-O 3 code(s)); (iii) disease locus (ICD topography code(s)); (iv) PubMed ID; (v) technique(s); (vi) series ID. For sample specific queries, additional features are available: sample ID; platform ID or description; and single or combined regional CNAs. Users can input gene name(s) in ''regional CAN'' search field. When at least two characters are entered into the field, suggestions based on a HUGO gene list are displayed for selection. Gene selections will be converted to genomic locations.

In the results table, associated array information is displayed. A number of links to additional and/or outside data is provided, according to the information available: the corresponding PubMed entries; the original GEO/AE accession display page for more complete information; the case and publication entries on the Progenetix website for further analysis; and importantly the array specific data visualization page.

Data download options. On pages resulting from sample queries or sample data processing, users are presented with options to download sample data based on the current queryÕs return. Currently, three different file types are offered: JSON files, tab separated feature files and segments list files. These files enable bioinformaticians to perform further analyses based on their tools of choice. Particularly, the JSON format can be used for direct database import (e.g. MongoDB) or can be deparsed by common libraries (e.g. JSON.pm), or being read into web applications.

Array probe data visualization. In the array plot interface, original plots of genomic array data sets can be searched and visualized (Figure S1). Default threshold parameters which were either provided with the data or assigned during the initial visualization will be loaded. In single array visualization, the general view of probe distribution and post-thresholding segmentation results are displayed for the whole genome as well as for each individual chromosome. If multiple arrays are retrieved, users can select sample data for downstream analysis procedures. Figure S2 shows the screenshot of single array visualization.

Users can segment the raw data values and re-plot the results after revising the following parameters:

- Golden path edition, default HG18/NCBI Build 36. This is still the commonly used version of the human reference genome assembly. At the moment, coordinates of probes from all platforms were remapped to HG18. For the near future, we intend to allow for a selection of updated genome editions.
- Chromosomes to plot, default 1 to 22. Single or all chromosomes can be selected for re-plotting. To avoid gender bias, most platforms do not contain probes in chromosome X and Y during the design.
- Loss/gain thresholds. Cut-offs from which a segment is considered a genomic loss or gain. The optimum thresholds may vary between platforms.
- Region size in kb. Sets a filter to remove CNA below (e.g. probable noise) or above (e.g. exclude non-focal CNA) a certain size range.
- Minimal probe numbers for segments. This parameter can be used to limit the minimal number of probes required for a segment to be considered (e.g. to remove aberrant segmentation due to probe level noise). Empirical examples would be values of 2–3 for high quality BAC arrays and 6–10 for Affymetrix SNP 6 arrays.
- Plot region. Single genomic region to be plotted, overriding the chromosome selection above. When selected, plots with this region will be generated for all current arrays. This is valuable to e.g. display the CNA status and copy number transition points for specific genes of interest (Figure S3).

Zoom-in visualization of focal CNA. Figure 2 shows the visualization of focal genomic imbalances, e.g. to identify genes of interest targeted by focal CNA. The whole genome view of GSM535547 (human high grade glioma sample analyzed by Agilent Human Genome CGH Microarray 244A) shows a small regional deletion in chromosome 9p21. When plotting the approximate locus of the deletion (specified as chr9:21600000-22400000), genes, probes and chromosome bands in this zoomed in region are shown. Two genes, MTAP and CDKN2A can be seen as being localized in a potential homozygously deleted region. The focal deletion of these known tumor suppressor genes [26,27] points to their specific involvement in the glioblastoma sample analyzed here.

Querying compound CNA. The concept of focal CNA detection can be integrated with a global search for arrays

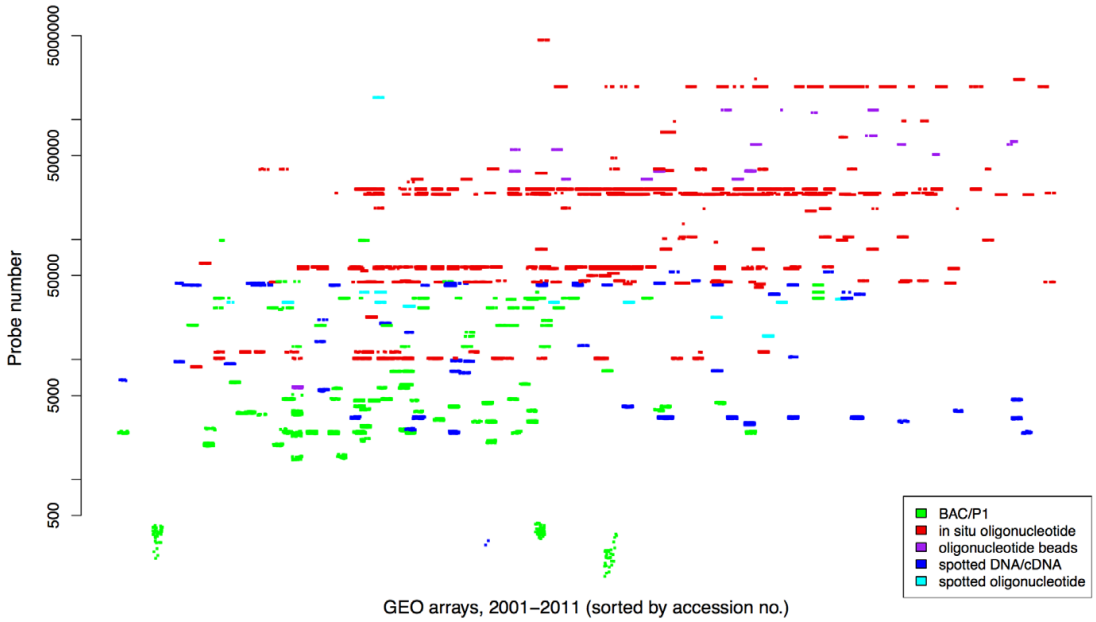


Figure 1. Distribution of resolutions and techniques of GEO platforms. Each point represents a genomic array. The Y axis is labeled with probe number in log scale. The X axis denotes the time sequence of array data generation. From left to right are years from 2001 to 2011.
doi:10.1371/journal.pone.0036944.g001

containing gene specific regional imbalances. As an example, we demonstrate the search for arrays displaying imbalances in 4 gene loci associated with glioblastoma: EGFR, a transmembrane receptor and proto-oncogene [28]; PTEN, a tumor suppressor gene [29]; ASPM, frequently overexpressed in glioblastoma relative to normal brain tissue [30]; and CDKN2A (see above). In the "Search Samples" form, the "Match (Multiptle) Regions & Types" can be used to specify the genomic regions of those four genes including the expected CNA type: for EGFR (chr7:55054219-55242524:1), PTEN (chr10:89613175-89718511:-1), ASPM (chr1:195319885-195382287:1) and CDKN2A (chr9:21957751-21984490:-1), respectively. When executing the query, these regions were matched with the whole database and returned cases which have imbalances overlapping all these regions. When excluding controls and "worst quality" datasets, 303 out of 42421 arrays could be identified matching all four CNA regions. In addition to glioblastoma, several other types of cancer cases were among the results, including e.g. neuroblastomas, breast carcinomas, melanomas and lung carcinomas, which is in accordance with some previous observations [31–34]. CNA and associated data of those cases can be processed by online tools for further analysis and visualization (Figure S4) or downloaded for offline processing.

Copy number profiling of selected cancer entities. One aim of arrayMap is to allow researchers to conveniently perform aCGH meta-analysis across different platforms. By selecting a single or several cancer entities e.g. based on their ICD entity codes or diagnostic keywords, users are able to generate disease specific CNA frequency profiles or to compare profiles of different cancer types.

As an example, we used ICD-O code 9440/3 (glioblastoma, NOS) to query the database. 1478 arrays from 25 publications were returned and passed to our suite of online analysis tools. Chromosomal ideograms and histograms were generated representing the frequency of copy number aberrations identified over the whole dataset (Figure 3A). In the overall aberration profile, the most common genomic imbalances included whole chromosome 7 gain and chromosome 10 loss, as well as focal gains e.g. on bands 1q21 and 17q21. In our example dataset, a prominent focal deletion hot-spot was centered around 9p21.3 (921 of 1478 arrays, 62.31%) which had been discussed previously [35]. The distribution of CNAs over the individual arrays was visualized through a matrix plot (Figure 3B). As additional information to the frequency histograms, this form of visualization facilitates e.g. the detection of CNA patterns among individual arrays as well as the concordance of individual CNAs (e.g. here the arm-level changes in chromosome 7 and 10).

In the matrix plot, clicking on a certain segment would open the related view in the UCSC genome browser [36], for detailed information related to this genomic region (SVG plot only). The plot order of arrays can be re-sorted according to ICD morphology, ICD topography, clinical group or PubMed ID, which can be helpful in associating CNA patterns to external classification categories. For the selected classification criterium (default: ICD morphology), regional CNA frequencies for cases matching the different values will be visualized through a heatmap (Figure 3C); this feature is especially useful when comparing a number of different primary classification criteria.

An Overall Genomic Copy Number Profile of Cancer

Our high quality core dataset in arrayMap was used to generate an overall cancer copy number aberration profile based on 29,137 arrays (Figure 4). This data represented 177 cancer types according to ICD-O 3 code, with 59 types among them contained

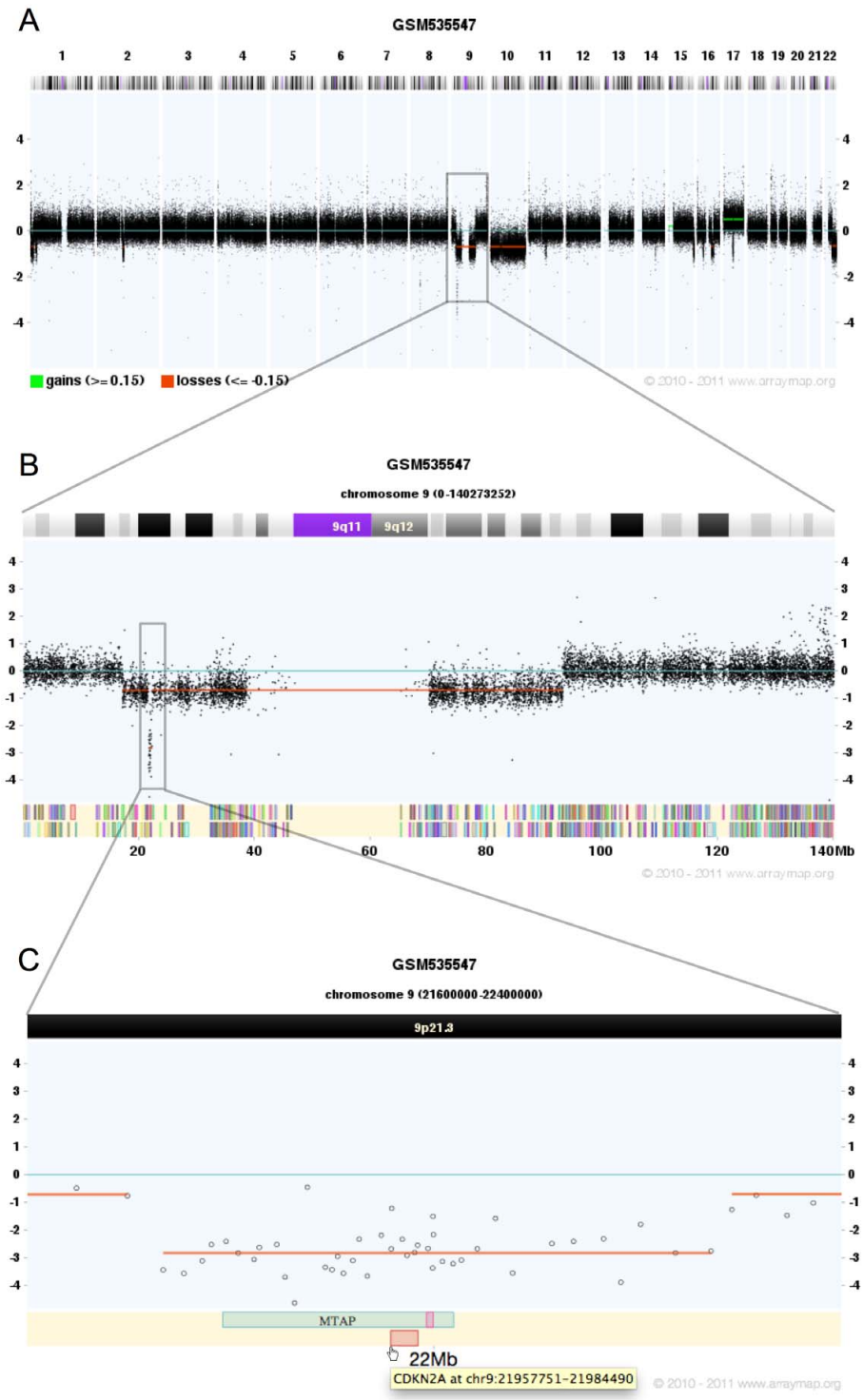

A
GSM535547
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
gains (>= 0.15)
losses (<= -0.15)
© 2010 - 2011 www.arraymap.org
B
GSM535547
chromosome 9 (0-140273252)
9q11
9q12
20 40 60 80 100 120 140Mb
© 2010 - 2011 www.arraymap.org
C
GSM535547
chromosome 9 (21600000-22400000)
9p21.3
MTAP
22Mb
CDKN2A at chr9:21957751-21984490
© 2010 - 2011 www.arraymap.org

Figure 2. Zoom-in visualization of focal CNA. (A) GSM535547 (human high grade glioma, Agilent CGH 244A) shows high quality of probe hybridization signal. CNAs are easy to distinguish. (B) When zoom-in the whole chromosome 9, an approximately 80 MB deletion is displayed, with two breakpoints located in p and q arm respectively. In addition, a small regional deletion in 9p21 is quite clear. Color bars in lower region of the panel represent 848 genes located in chromosome 9. (C) Zoom in the potential homozygously deleted region in 9p21 by specifying the exact region: chr9:21600000-22400000. The zoomed-in plot shows probes, chromosome band and two tumor suppressor genes, MTAP and CDKN2A. Gene name and location will be given while mouse hover. They link to UCSC genome browser with additional information.
doi:10.1371/journal.pone.0036944.g002

more than 50 arrays. Overall, one of the most common genomic alteration is copy-number gain of chromosome band 8q24, which is found in 30% of total samples. According to the COSMIC [37] database, the most significant cancer gene in this region is MYC. It is a well-documented oncogene codes for a transcription factor that is believed to regulate the expression of 15% of all genes, including genes involved in cell division, growth, and apoptosis [38,39]. Other common imbalances observed in at least 25% of oncogenomic arrays included gains of regions on e.g. 17q21 (29%), 1q21 (33%) and loss of regions on e.g. 8p23 (32%) and 9p21 (25%), including focal deletions of the CDKN2A/B locus (Figure 2).

While the overall CNA frequency distribution points towards DNA features targeted in multiple entities, this information is insufficient for deriving molecular mechanisms associated with specific cancer types. The genomic heterogeneity of different neoplasias is reflected in the varying patterns of regional CNA frequencies. Based on our core dataset, we have generated a heatmap-style visualization of frequency profiles for all ICD-O entities containing more than 50 arrays (Figure S5). The striking

Figure 3. Copy number profiling of glioblastoma. (A) Chromosomal ideogram and histogram showing frequency of copy number aberrations. Percentage values corresponding to gains (yellow) and losses (blue) identified over the whole dataset. The most frequent imbalances include gain of chromosome 7 and loss of chromosome 10, 9p21.3. (B) Matrix plot of 1478 glioblastoma cases. The Y axis represents individual samples. The distribution of genomic copy number imbalances reveals the individual aberration patterns of glioblastoma. (C) Heatmap of regional CNA frequencies for 1478 arrays. The intensity of green and red color components correlates to the relative gain and loss frequencies, respectively. If dataset contains cancer subtypes, cancers with similar CNA frequency profiles will be clustered together, such that differences between subtypes will be revealed (e.g. see Figure S4H).
doi:10.1371/journal.pone.0036944.g003

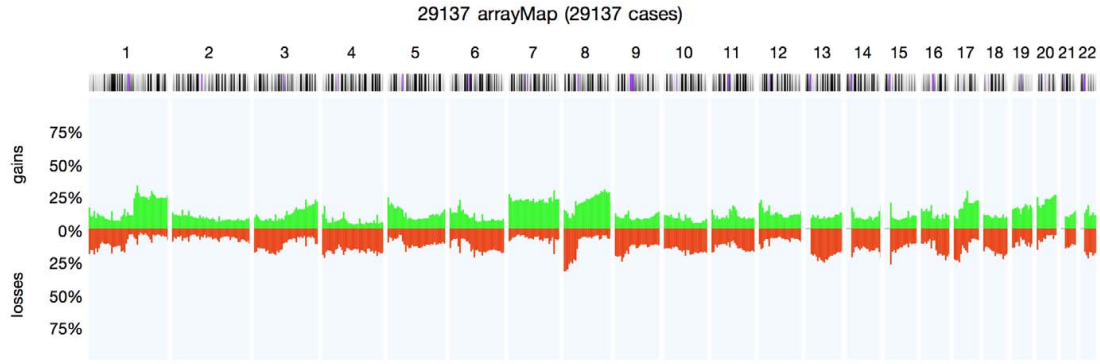


Figure 4. The overall cancer copy number aberration profile consisted of 29137 arrays. This plot represents 177 cancer types according to ICD-O 3 code. Percentage values in Y axis corresponding to numbers of gains (green) and losses (red) account for the whole dataset.
doi:10.1371/journal.pone.0036944.g004

patterning of the CNA profiles indicates the non-random occurrence of CNAs, and should be seen as an invitation to explore e.g. CNA similarities shared by separate histopathological entities, as a way to transpose knowledge about pathophysiological mechanisms.

Discussion

arrayMap was developed to facilitate the progress of oncogenomic research. Our aim is to provide high-quality genomic copy number profiles of human tumors, along with a set of tools for accessing and analyzing CNA data. The service has been implemented with a straightforward web interface, including search options for CNA features and clinical annotation data. All assembled datasets are processed into platform independent segmentation and, for the vast majority of arrays, probe level data files, and are presented in consistent formats. Importantly, the direct access to precomputed probe level data plots supports a rapid evaluation of experiments for features of interest. As a curated database using standardized annotation schemes (e.g. ICD classification), arrayMap facilitates the exploration of cancer type specific CNA data, as well as the statistical association of genomic features to clinical parameters.

arrayMap is a dynamic database that is being continuously expanded and improved. We will review the existing and newly published articles to update the database periodically. Over the past decade, we have witnessed a rapidly increasing number of aCGH publications, which gives us sufficient evidences to anticipate that cases in our database will continue to be deposited at a high rate. Although arrayMap is not a user driven repository, we welcome and support users interested in using the site for yet undisclosed data, if they agree on data sharing upon publication.

Although, in contrast to the continuous data from expression analysis, copy number analysis explores discrete value spaces (countable number of DNA copies, for segments defined by genomic base positions), interpretation of the data can vary due to different low level (e.g. signal/background correction) and higher level (e.g. segmentation algorithms, regional or size based filtering) procedures. In that respect, we have to emphasize that the results of our data processing and annotation procedures are open to scrutiny. We encourage a critical review of individual results, and are open for suggestions regarding improved processing procedures for specific platforms.

In this paper, we have provided example scenarios of using arrayMap on different levels, i.e. locus centric and for entity profiling. We believe that systematic analyses will help researchers to discover features which are indiscernible in individual studies, and thus bring new insights for understanding of disease pathology and the development of new therapeutic approaches [40–43]. We expect that researchers will integrate arrayMap data with their own analysis efforts, e.g. to increase sample size or for result verification purposes. We hope that this database will promote further evolution of microarray data meta-analysis. ArrayMap provides access to more than 200 tumor types, which makes it suitable for research across cancer entities. Furthermore, normal sample controls are of vital importance for genomic imbalances studies. ArrayMap includes more than 3000 normal samples from healthy individuals or from normal tissues of cancer patients. These data could be integrated as reference dataset e.g. to account for copy number variation data superimposed on the tumor profiling results.

In the near future, with the continuous accumulation of very high resolution CNA data from genomic arrays and next-generation sequencing experiments, it will become possible to integrate these data into systems biology methods to elucidate effects of genomic instability, and describe the results from more perspectives. Envisioned examples would be e.g. the identification of genes that are involved in metastasis and treatment response; identification of chromosomal breakpoints distribution in cancer; and modeling functional networks in cancer by systems biology approaches.

Methods

Dataset Collection

Raw experimental data from a variety of platforms and repositories were extracted. They were converted to an uniform format which is suited to our reanalysis and visualization system. After a series of parsing procedures, the called copy number data is stored in arrayMap. The flowchart of arrayMap data collection and analysis is as shown in Figure 5. Five main data sources are integrated into arrayMap:

GEO/AE. For extracting appropriate data Series from GEO/AE, two basic criteria have to be fulfilled. First, the raw data has to be from human malignancies analyzed by BAC, cDNA, aCGH or oligonucleotide arrays. Second, the array platform must be genome wide, with the optional omission of the sex chromosomes.

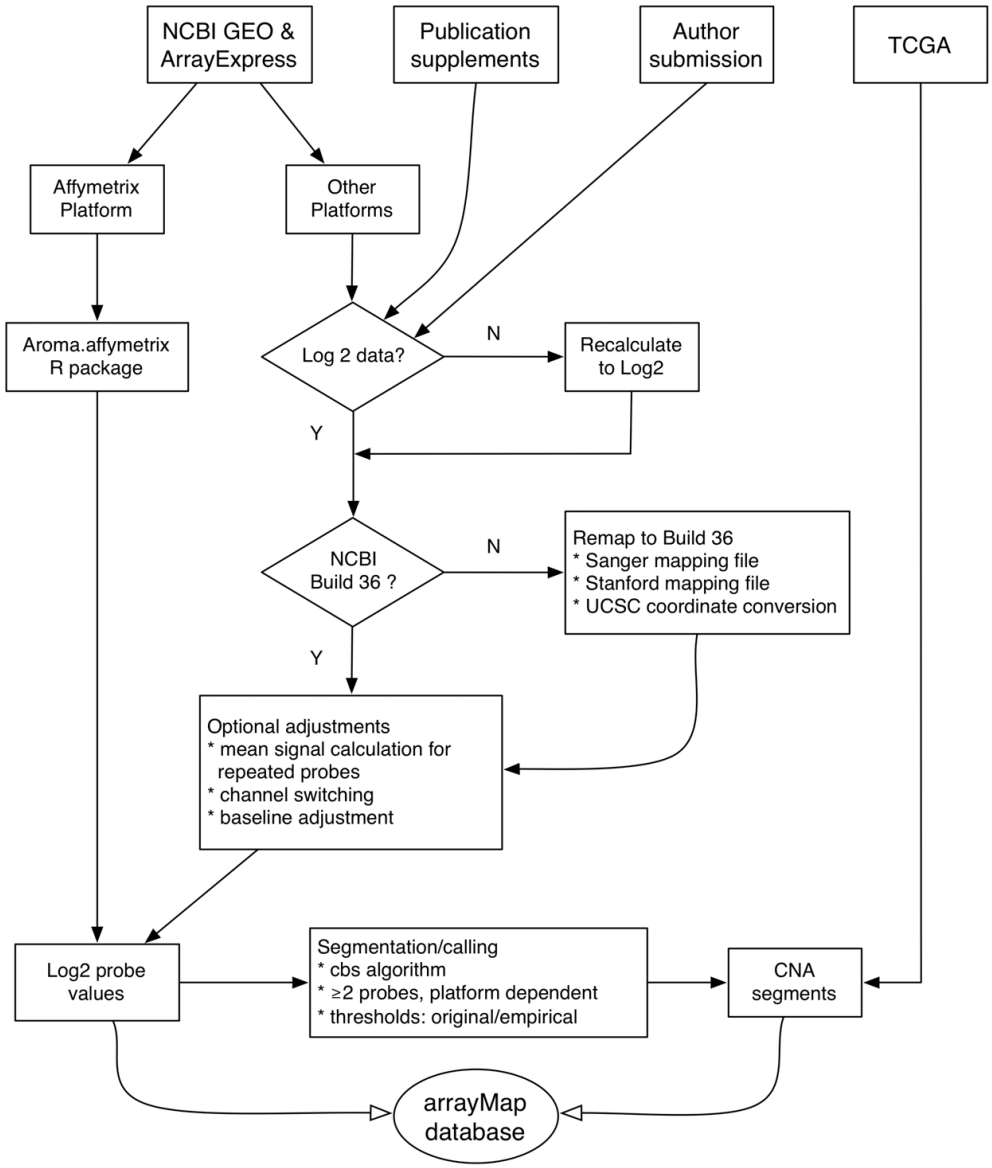


Figure 5. The flowchart of arrayMap data collection and analysis procedures. Publicly available raw data or segmented data was collected from the respective data sources. Files were re-processed by distinct procedures, according to the different data types. Probe coordinates were remapped to the most commonly encountered human reference genome assembly (NCBI Build 36/hg18). All probe specific ratios were converted to log2 values. Thresholds for genomic gain and loss were obtained from the original publications or series annotations; if not available, empirical thresholds were assigned. A minimum of 2 probes was required for calling a CNA segment, with higher values used on high-density arrays and/or in cases of excessive probe level noise. Processed probe and segment information was converted to uniform formats and stored in per-sample text files, which are accessed through the arrayMap web applications.
doi:10.1371/journal.pone.0036944.g005

Chromosome or region specific arrays were excluded because they were not able to reveal the whole genomic profile of the respective cancer. Associated clinical data was extracted if available.

TCGA. Segmentation data with available clinical information was extracted and incorporated into the database. Due to data sharing restrictions, TCGA data is an exception in that, so far no probe level data is incorporated into arrayMap. This exception

was accepted since users will be able to access individual TCGA datasets through the projects web portal at http://tcga-data.nci.nih.gov/tcga/.

Publications. Many aCGH datasets can be found in the text or supplementary files of publications. In order to collect data from publications, we relied on our Progenetix projectÕs setup. Data in Progenetix is manually curated. The collection strategies are:

- literature mining using complex search parameters through PubMed
- identification of called aCGH data, in GP annotation or tabular format (article, supplementary tables)
- evaluation of supplementary files for probe specific data tables
- follow-up on article links outs, to repository entries or referenced datasets

User submission. User submitted data was provided in a number of formats which were converted to the standard format as described. Although we accept and support private datasets, we insist on integration of at least the genomic and core clinical data (e.g. disease classifiers) upon publication of the datasets analysis results.

Dataset Analysis

Probe remapping. A pipeline has been generated for determining the genomic positions for the tens to hundreds of thousands array probes with reference to a common genome Golden Path edition. For each array platform, the genome positions of probes were remapped to the current commonly used version of the human reference genome assembly (NCBI Build 36.1/hg18). Specific mapping procedures were employed for different types of probes. BAC clones were firstly remapped according to the clone sets information of Sanger/DECIPHER database [44]. If the probe position was not available, the UCSC Genome annotation database [36] (release hg18) was used for compensation. After these two steps, a mean of 98% of the BAC clones were remapped. For IMAGE clone sets, only the UCSC Genome annotation database was used. The average remapping rate of IMAGE clones was 91%. Affymetrix raw CEL data files were analyzed based on hg18 library files, namely the output segments have hg18 coordinates. The summary of the percentage of mapped probes is given in Table 3. The mapping details for each platform can be found in the (Table S4).

Probe signal normalization. The array data available was given in a variety of formats, most frequently as log2 ratio of probe hybridization intensity. In order to make data from different platforms directly comparable, all other types of normalized values were converted to log2. For dye swap experiments, reference/tumor intensity ratios data was "reversed" representing a tumor/reference value. For some two-color arrays for which only raw signal intensity were provided, the normalized log2 ratio for each probe was calculated by.

$$r = \log 2((T_s - T_b)/R_s - R_b).$$

where T_s and T_b represent tumor sample intensity and tumor channel background intensity respectively, and R_s and R_b represent reference sample intensity and reference channel background intensity respectively. If multiple instances of the same clone exist, the average signal intensity of the certain clone was considered.

To call gains and losses according to normalized log2 ratio is an important step to identify copy number imbalances. For each re-analyzable dataset, related publications were explored to obtain original threshold descriptions. If this information was not available, empirical thresholds were assigned and resulting CNA calls were visually compared with probe value plots. Processing method and threshold information for each array are provided in the Table S5.

Affymetrix genotyping arrays. For the widely used Affymetrix GenomeWide SNP arrays, raw CEL files were downloaded and underwent a massive re-analysis using the R package aroma.affymetrix [45] with the CRMAv.2 method [46]. During the processing step, approximately 50 normal sample arrays were employed as a reference set for each array type to reduce the noise level. Normal tissue arrays from different labs were extracted and used to build the reference dataset. In order to obtain high quality arrays, we excluded arrays which contain segments greater than 3 mega-bases, since copy number variations are always smaller than 3 mega-bases. The list of normal tissue reference arrays is giving in Table S6.

Quality control. In our review of array data deposited in GEO or collected from publication supplements we encountered a large number of individual data sets with insufficient or limited probe quality. Also, for samples of unprocessed raw data (e.g. Affymetrix CEL files), we found that QC measures reported previously (e.g. call rate [47], NUSE [48], RLE [48]) only had a limited accuracy for detection of arrays with inadequate probe level data. Currently, the most viable strategy for quality assessment of processed, heterogeneous copy number arrays is the visual inspection of probe plotting and segmentation results through an experienced researcher. For the first arrayMap edition we generated a quality classification system, which contains a total of 4 categories based on inspections of genome-wide array plots:

- Excellent. Probe signal distribution is significantly different between normal regions and imbalance regions. Signal baseline is distinct and unique, making segmentation threshold realistic appearing. Chromosomal changes are pretty clear.
- Good. In general good quality. Probe signal may contain some noise, but tolerable. Chromosomal changes are distinguishable.

Table 3. Percentage of remapped probes according to platform types.

Platform type	Average mapping rate	Number of arrays	Number of GPLs
Original HG18 (Build 36)	NA	1583	40
in situ oligonucleotide	99%	21678	55
BAC/P1	98%	5464	55
spotted DNA/cDNA	91%	2365	82

doi:10.1371/journal.pone.0036944.t003

- Hypersegmented. Serrated distribution of probe signal intensities, causing dozens of separate peaks and discontinuous segments. Chromosomal changes are always up to several hundreds and smaller than 5 mega-bases.
- Noisy. Probe signal intensities are highly scattered, but well-distributed, with high standard deviation, resulting in the inability to differentiate copy number changes.

Depending on the intended research purpose this basic classification system can be used for a pre-analysis triage of copy number data. Applying stringent review criteria we identified a core dataset with "excellent" quality arrays accounting for approximately 60 percent of total arrays. We are currently working on a platform independent quality assessment system for genomic arrays, which will be implemented in future versions of the arrayMap resource.

Associated data. For arrayMap, data is stored with separate datasets for each array. This is in contrast to the Progenetix database, for which technical replicates where available are combined into case specific CNA profiles. In arrayMap, technical replicates are assigned an identical case identifier to facilitate downstream statistical procedures including e.g. clinical data correlations. The assignment of the correct diagnostic entity to each sample is an essential step in generating a binding between genomic and associated data points. At the same time, to ensure annotation consistency and make the retrieval process more efficient, for all CNA profiles the following data points were manually collected from GEO/ArrayExpress and published papers if available.

- Descriptive diagnostic text, as available through the original source
- Diagnostic classification according to the International Classification of Diseases in Oncology (ICDO 3, morphology with code)
- Tumor locus according to ICD (ICD topography with code)
- Source of material (e.g. primary tumor, cell line, metastasis)
- Clinical parameters where available, including age, gender, grade, clinical stage (TNM coded), recurrence/progression, time to recurrence/progression, death and followup

Web Server. An online interface of arrayMap database was created using Perl common gateway interface (CGI) and R scripts running on Mac OS X Server. Sample and series data is stored using a MongoDB database eingine (http://www.mongodb.org). Precomputed array plots are stored as flat files, mostly in both SVG and PNG versions. The online release of the service has been optimized to be compatible with major browsers supporting current web standards (CSS2, HTML5, XML with inline SVG; e.g. Safari $>=3.0$, Firefox $>=3.0$, InternetExplorer $>=9$, Google Chrome) with limited fallback support. Dynamic graphics provided in the array plot module were implemented as server side services by technologies including XML/XHTML, JavaScript, SVG and HTML5 Canvas.

For the future, we intend a quarterly database content revision to ensure inclusion of newly published articles and GEO/AE entries. Archived versions of the sample annotations will be made available upon special request. Additional feature and small data updates will be performed as seen necessary. The "News" page of Progenetix/arrayMap will be used for feature and content announcements.

Supporting Information

Figure S1 Array data sets visualization. Original plots and optimized parameters for GSE21530 which contains 8 intimal sarcoma samples hybridized on Agilent CGH Microarray 244A platform. The normalized probe signal log2 ratios and post-thresholding segmentation results for each array are intuitively displayed. Genomic alterations are represented by horizontal green (gain) and red (loss) lines. Alterations defined here as regions with log2 ratio >0.15 or <-0.15. Simplified schemas of CNAs link to UCSC genome browser for further review.
(PDF)

Figure S2 Screenshot of single array visualization. ArrayMap plots for GSM630977 (acute myelogenous leukemia). Besides the whole genome view, subviews of each chromosome are displayed as well. From these plots, different kinds of genetic variation events are clearly revealed, e.g. massive genomic rearrangement in chromosome 6; arm-level gain of chromosome 8q and 3MB focal change around 1p31.3. Through the "Plot Array Data" interface, users can segment the raw data values and re-plot the results with customized parameters.
(PDF)

Figure S3 Plot single genomic region. In the "Plot Array Data" interface, input the precise location (chr5:1100000-1400000) in "Plot Region" field. Plots with this region were generated for all 8 arrays in the current series (GSE21530). In this region, there are 5 genes which are shown schematically as colored boxes. CNA status and copy number transition points for these genes are displayed.
(PDF)

Figure S4 Compound CNA query. (A) Four gene loci associated with glioblastoma (EGFR, PTEN, ASPM and CDKN2A) were inserted into "Match (Multiple) Regions & Types" field. 303 out of 42421 arrays were returned. (B) Classification information of these 303 arrays were displayed and can be selected for the following analysis. (C) Statistical and plot parameters can be customized. Associated data was processed by online tools, and returned results included: (D) Chromosomal ideogram and (E) histogram, show frequency of copy number aberrations; (F) Matrix plot reveals the aberration pattern of selected arrays; (G) Array classification tree generated by hierarchical Ward clustering, arrays with similar frequency of CNA are part of the tree branch. (H) Heatmap of CNA frequencies clustered by clinical group.
(PDF)

Figure S5 Heatmap of frequency profiles for 59 cancer types. Heatmap visualization of frequency profiles for all ICD-O entities containing more than 50 arrays in our core dataset. Region specific gain/loss frequencies were mapped to 1MB intervals. The intensity of colors (green: gains; losses: red) corresponds to the relative frequency of CNAs for each interval.
(PDF)

Table S1 Entities extracted from NCBI GEO and EBI ArrayExpress.
(XLS)

Table S2 Cancer entities grouped by ICD-O code.
(XLS)

Table S3 Platform type distribution in arrayMap.
(XLS)

Table S4 Probe remapping rate for platforms.
(XLS)

Table S5 Processing method and threshold for calling genomic gains and losses.
(XLS)

Table S6 Normal tissue reference arrays for Affymetrix platforms.
(XLS)

Acknowledgments

We want to thank Christian von Mering, Homayoun Bagheri, Henrik Bengtsson and Nuria Lopez-Bigas for helpful discussions.

Author Contributions

Conceived and designed the experiments: HC NK MB. Performed the experiments: HC MB. Analyzed the data: HC NK MB. Contributed reagents/materials/analysis tools: HC NK MB. Wrote the paper: HC MB.

References

1. Stallings RL (2007) Are chromosomal imbalances important in cancer? Trends in genetics : TIG 23: 278–283.
2. Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, et al. (2006) DNA copy number amplification profiling of human neoplasms. Oncogene 25: 7324–7332.
3. Weir BA, Woo MS, Getz G, Perner S, Ding L, et al. (2007) Characterizing the cancer genome in lung adenocarcinoma. Nature 450: 893–898.
4. Wiedemeyer R, Brennan C, Heffernan TP, Xiao Y, Mahoney J, et al. (2008) Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. Cancer cell 13: 355–364.
5. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, et al. (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 446: 758–764.
6. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068.
7. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466: 869–873.
8. Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, et al. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and geneexpression subtypes of breast cancer. Genes, chromosomes & cancer 45: 1033–1040.
9. Hu X, Stern HM, Ge L, O'Brien C, Haydu L, et al. (2009) Genetic alterations and oncogenic pathways associated with breast cancer subtypes. Molecular cancer research : MCR 7: 511–522.
10. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science (New York, NY) 258: 818–821.
11. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature genetics 23: 41–46.
12. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome research 14: 287–295.
13. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative metaanalysis of chromosomal CGH data. BMC cancer 7: 226.
14. Alloza E, Al-Shahrour F, Cigudosa JC, Dopazo J (2011) A large scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression. BMC medical genomics 4: 37.
15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic acids research 39: D1005–10.
16. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, et al. (2010) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research 39: D1002–D1004.
17. Scheinin I, Myllykangas S, Borze I, Böhling T, Knuutila S, et al. (2008) CanGEM: mining gene copy number changes in cancer. Nucleic acids research 36: D830–5.
18. Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, et al. (2011) CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic acids research 39: D968–74.
19. Baudis M, Cleary ML (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics (Oxford, England) 17: 1228–1229.
20. Baumbusch LO, Aarøe J, Johansen FE, Hicks J, Sun H, et al. (2008) Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. BMC genomics 9: 379.
21. Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, et al. (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC genomics 10: 588.
22. Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, et al. (2007) A comparison of DNA copy number profiling platforms. Cancer research 67: 10173–10180.
23. Bengtsson H, Ray A, Spellman P, Speed TP (2009) A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. Bioinformatics (Oxford, England) 25: 861–867.
24. Heinrichs S, Look T (2007) Identification of structural aberrations in cancer by SNP array analysis. Genome biology. pp 1–5.
25. Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics 39: S16–S21.
26. Lubin M, Lubin A (2009) Selective killing of tumors deficient in methylthioadenosine phosphorylase: a novel strategy. PloS one 4: e5735.
27. Krasinskas AM, Bartlett DL, Cieply K, Dacic S (2010) CDKN2A and MTAP deletions in peritoneal mesotheliomas are correlated with loss of p16 protein expression and poor survival. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 23: 531–538.
28. Smith JS, Tachibana I, Passe SM, Huntley BK, Borell TJ, et al. (2001) PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. Journal of the National Cancer Institute 93: 1246–1256.
29. Li J (1997) PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast, and Prostate Cancer. Science (New York, NY) 275: 1943–1947.
30. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, et al. (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences of the United States of America 103: 17402–17407.
31. Zhang W, Zhu J, Bai J, Jiang H, Liu F, et al. (2010) Comparison of the inhibitory effects of three transcriptional variants of CDKN2A in human lung cancer cell line A549. Journal of experimental & clinical cancer research : CR 29: 74.
32. van der Rhee JI, Krijnen P, Gruis NA, de Snoo FA, Vasen HFA, et al. (2011) Clinical and histologic characteristics of malignant melanoma in families with a germline mutation in CDKN2A. Journal of the American Academy of Dermatology.
33. Bourdeaut F, Isidor B, Ferrand S, Thomas C, Moreau A, et al. (2011) Homozygous PTEN deletion in neuroblastoma arising in a child with Cowden syndrome. American journal of medical genetics Part A 155: 1763–1766.
34. Jin K, Kong X, Shah T, Penet MF, Wildes F, et al. (2011) Breast Cancer Special Feature: The HOXB7 protein renders breast cancer cells resistant to tamoxifen through activation of the EGFR pathway. Proceedings of the National Academy of Sciences of the United States of America.
35. Wiltshire RN, Rasheed BK, Friedman HS, Friedman AH, Bigner SH (2000) Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification. Neurooncology 2: 164–173.
36. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic acids research 39: D876–82.
37. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research 39: D945–50.
38. Gearhart J, Pashos EE, Prasad MK (2007) Pluripotency redux–advances in stem-cell research. The New England journal of medicine 357: 1469–1472.
39. Dalla-Favera R, Bregni M, Erikson J, Patterson D, Gallo RC, et al. (1982) Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proceedings of the National Academy of Sciences of the United States of America Vol. 79: 7824–7827.
40. Climent J, Dimitrow P, Fridlyand J, Palacios J, Siebert R, et al. (2007) Deletion of chromosome 11q predicts response to anthracycline-based chemotherapy in early breast cancer. Cancer research 67: 818–826.
41. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, et al. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer cell 10: 529–541.
42. Stevens KN, Fredericksen Z, Vachon CM, Wang X, Margolin S, et al. (2012) 19p13.1 is a triple negative-specific breast cancer susceptibility locus. Cancer research.
43. Park NI, Rogan PK, Tarnowski HE, Knoll JHM (2012) Structural and genic characterization of stable genomic regions in breast cancer: Relevance to chemotherapy. Molecular oncology.
44. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, et al. (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics 84: 524–533.

45. Bengtsson H, Simpson K, Bullard J, Hansen K (2008) aroma.affymetrix: A genetic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Tech Report #745 Department of Statistics, University of California, Berkeley.
46. Bengtsson H, Wirapati P, Speed TP (2009) A single-array preprocessing method for estimating fullresolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics (Oxford, England) 25: 2149–2156.
47. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, et al. (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology 34: 591–602.
48. F C, AL A, SA K, TP S, VL SM (2005) NUSE and RLE: Quality assessment of oligonucleotide microarray data to quantify systemic variation. 2005 Meeting of the Federation of Clinical Immunology Societies Boston, MA.

5 THE RELATIONSHIP BETWEEN COPY NUMBER ABERRATIONS AND FUNDAMENTAL GENOME STRUCTURE

5.1 PREFACE

DNA copy number aberrations are found in almost all solid tumors and contribute to tumor progression[3]. The segment size may range from the entire chromosome arm to less than 100 kb. Recent cytogenetic studies revealed the heterogeneity of neoplasia-associated chromosomal aberrations[3]. There is a hypothesis that the variable extent of CNA would tend to indicate different pathways taken by individual tumors during the course of their genetic evolution[45]. In this way, the profiling of copy number alterations is likely to reflect a series of non-neutral selection procedures in cancer genome. Several studies have performed large-scale aCGH data analyses on a wide range of cancer types, focused on detecting “driver” mutations from recurrent CNAs, which may point to tumor suppressors and oncogenes[82]. However, little is known about specific gene defects that give rise to chromosomal alterations in tumors. CNA hotspots encompass hundreds of gene loci, far beyond the involvement of a limited set of strong “cancer driver” genes. Moreover, even such “regions of interest” are usually highly discordant in different studies of the same tumor type. Apart from recurrent CNAs, sporadic gains and losses can occur throughout the genome. As a result of genomic instability, these changes may also have an effect on tumor progression, rather than just “byproducts”.

As we know, human genome has a complex structure, and it is exemplified by the non-uniform distribution of genes, both within and between chromosomes[194]. The clustering of genes in the genome, known as gene-rich region, has been taken into account in the array probe design. The fundamental aspects of human genome organization are likely to have important implications for understanding the non-neutral selection of CNA regions. However, the relationship between CNAs and the pattern of gene distribution has so far not been systematically investigated. The rapid accumulation of oncogenomic data in our arrayMap database enables me to examine this question in detail.

5.2 CNA ENRICHED IN GENE-RICH REGIONS

For this research project, to explore the possible correlation between CNAs and local gene density across the genome, I collected 16,264 copy number profiles of 62 diagnostic groups from arrayMap. These data represent 19,471 high quality arrays, and were collected from 3 public resources, including GEO, ArrayExpress and supplemental materials of publications. Among the arrays, 14,837 samples came from 28 diagnostic groups, each represented by more than 100 specimens. There are a total of 298,904 gains and 307,414 losses, for an average of 18 gains and 19 losses per sample.

In my strategy to investigate the possible non-neutral selection for CNAs in cancer, the genome was firstly divided into 10 Mb non-overlapping intervals, then the number of genes within each interval was counted. The definition of gene-rich region here is the genomic intervals that contain twice as many as the average 73 genes in 10 Mb intervals. Subsequently, I determined the number of samples that exhibit CNAs in each interval, and considered genomic gains and losses respectively. Since many CNAs encompass thousands of gene loci, far beyond the size of interval, only focal CNAs were considered, which were defined here as being smaller than 5 Mb. I found that focal genomic gains were significantly positively correlated with gene-rich regions. To evaluate whether the observed correlation could be obtained by chance, I randomly shuffled the number of genes in intervals across the genome, and calculated the correlation coefficient. I repeated this permutation 10,000 times and the *p*-value is significant. I also observed a positive correlation between focal genomic losses and gene-rich regions.

To determine the correlation at different interval sizes (5 Mb and 15 Mb) and CNA sizes (smaller than 2 Mb and 10 Mb), I followed the method described above, i.e. I identified 5 Mb (or 15 Mb) non-overlapping intervals tiling the genome. For each such interval, I counted the number of genes and samples present focal CNAs (smaller than 2 Mb or 10 Mb). For each combination of these parameters, I found that focal CNAs were significantly enriched in gene-rich regions. To explore whether this positive correlation was only driven by cancer genes, I downloaded the catalogue of cancer gene from COSMIC database and excluded all intervals that contain cancer genes. After this step, the positive correlation was still observed. This suggests that the correlation is probably genuine.

For the correlation between CNA breakpoints and gene density, I again divided the genome into 1 Mb non-overlapping intervals and counted the number of breakpoints within each interval. The breakpoints, like CNAs, were also organized in hotspots across the genome. I found that breakpoints were positively correlated with gene-rich regions. Furthermore, this observation was independent of the size of intervals. These data indicated that the highly variable genomic regions were largely correlated with gene density across the entire genome.

In order to control the robustness of these results, several robustness checks were performed. For example, centromeres and telomeres are often substrates for rearrangements that are associated with structural genomic alterations in cancer and were excluded. Moreover, centromeres and telomeres are also common outliers of CNA breakpoints. For breakpoint analysis, I excluded all 1 Mb intervals that extended to centromeres and telomeres of each chromosome. As expected, CNA breakpoints were enriched in gene-rich regions after robustness control steps. Since the input data set is composed of genomic profiles from 180 array platforms, I divided these platforms into 3 groups according to their probe numbers and techniques (BAC/P1 and DNA/cDNA, oligonucleotide ≧ 200K and oligonucleotide < 200K). I repeated the correlation analysis on each of these platform groups, and found the same trend: the focal CNAs were significantly enriched in gene-rich regions. Therefore, the conclusions are not biased by the resolution of platforms. As for cancer types, cancer type specific analysis were performed. Six clinical groups that each has more than 650 samples were selected and all indicated a similar positive correlation between focal CNAs and gene-rich region. Thus, these observations were free of bias by using data from multiple cancer types. The non-neutral selection of small CNAs may be a recurring feature of cancer genome evolution. Overall, these results present the landscape of copy number changes in cancer, as well as its correlation with the fundamental genome structure, promote a basic understanding of human cancer.

The manuscript is included below.

The landscape of cancer genomes reveals correlation between somatic copy number aberrations and fundamental genome structure

Haoyang Cai[1,2,#], Nitin Kumar[1,2,#], Ni Ai[1,2], Christian von Mering[1,2], Michael Baudis[1,2*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

[2]Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

Abstract

Copy number aberrations (CNAs) are a hallmark of cancer genomes. Recurrent CNA hotspots are known to target well defined cancer genes. However, the average extent of CNA encompasses thousands of gene loci, far beyond the involvement of a limited set of strong “cancer driver” genes. To explore indications for a non-neutral selection for CNA regions, we analyzed a possible relationship between CNA occurrence and local gene density across the genome. We performed a genome-wide analysis of 606,318 somatic CNAs from 16,264 cancer samples, classified into 62 diagnostic groups. We observed a significant enrichment for small and extended localized CNA in gene-rich regions, both in terms of gene numbers and coding regions. The 1,212,636 DNA breakpoints that were associated with the CNA also showed a positive correlation with gene-rich regions. However, the frequency of arm-level CNA was negatively correlated with gene numbers on chromosome arms. Notably, when investigating cancer and platform type specific data, we got similar results. The landscape of cancer genomes revealed a global correlation between CNA and genome structure, and provides support for a non-neutral selection of multiple and/or extended CNA as recurring feature of cancer genome evolution.

Introduction

Genomic DNA copy number aberrations are commonly found in solid tumors, and contribute to tumor genesis and development. The size of these alterations may range from entire chromosome arm to less than 100 kb. The heterogeneity of neoplasia-associated chromosomal aberrations was comprehensively demonstrated by cytogenetic studies. The variable extent of CNA would tend to indicate different pathways taken by individual tumors during the course of their genetic evolution. Thus, the profiling of copy number alterations is likely to reflect a series of non-neutral selection procedures in cancer genome.

During the last decade, array comparative genomic hybridization (aCGH) and SNP arrays have been extensively employed to detect genome-wide copy number changes in cancer. The value of these techniques is further illustrated by recent studies in which CNAs were used for tumor subtype classification and cancer gene screening. In particular, several studies performed large-scale aCGH data analyses on a wide range of cancer types, provided striking new insights into human cancers. So far, many efforts have been focused on detecting “driver” mutations from recurrent CNAs, which may point to tumor suppressors and oncogenes. However, little is known about specific gene defects that give rise to chromosomal alterations in tumors. CNA hotspots encompass hundreds of gene loci, far beyond the involvement of a limited set of strong “cancer driver” genes. Moreover, even such “regions of interest” are usually highly discordant in different studies of the same tumor type. Apart from recurrent CNAs, sporadic gains and losses can occur throughout the genome. As a result of genomic instability, these changes may also have an effect on tumor progression, rather than just “byproducts”.

The complexity of human genome structure is exemplified by the non-uniform distribution of genes, both within and between chromosomes. The clustering of genes in the genome, known as gene-rich region, has been taken into account in the array probe design. The fundamental aspects of human genome organization are likely to have important implications for understanding the non-neutral selection of CNA regions. However, the relationship between CNAs and the pattern of gene distribution has so far not been systematically investigated. The rapid accumulation of oncogenomic data resulted in the

development of several online resources. These data enable us to examine the above question in detail.

In this study, to explore the possible correlation between CNAs and local gene density across the genome, we collected 16,264 publicly available copy number profiles of 62 diagnostic groups. We observed that small and extended localized CNAs were significantly enriched in gene-rich regions, either in terms of gene numbers or the fraction of coding regions. Furthermore, DNA breakpoints that occurred during chromosomal rearrangement also showed enrichment over gene-rich regions. Notably, these observations were free of bias by using data from multiple cancer types and array platforms. We propose that the non-neutral selection of small CNAs is a recurring feature of cancer genome evolution. Overall, our results present the landscape of copy number changes in cancer, as well as its correlation with the fundamental genome structure, promote a basic understanding of human cancer.

Results

A collection of copy number profiles from 19,471 cancer genomes

Our cancer copy number profiles were obtained from arrayMap database, a repository for annotated oncogenomic array data. The 19,471 high quality arrays were collected from 3 public resources, including NCBI Gene Expression Omnibus (GEO), EBI ArrayExpress and supplemental materials of publications, mostly from GEO. For the whole data set, ~97% samples are primary tumors, with the remainder cell lines, metastasis or relapse. After removing technical repeats (e.g. different platforms for one sample), a total of 16,264 samples from 390 studies were left. Among the rest, 14,837 samples came from 28 diagnostic groups, each represented by more than 100 specimens. The summary and details of input data set are listed in **Supplementary Table 1 and 2**, respectively. In our data set, we observed a total of 298,904 gains and 307,414 losses, for an average of 18 gains and 19 losses per sample. In general, hematological malignancies showed less CNAs than epithelial carcinomas. The distribution of sample and CNA number across major cancer types are shown in **Figure 1a**.

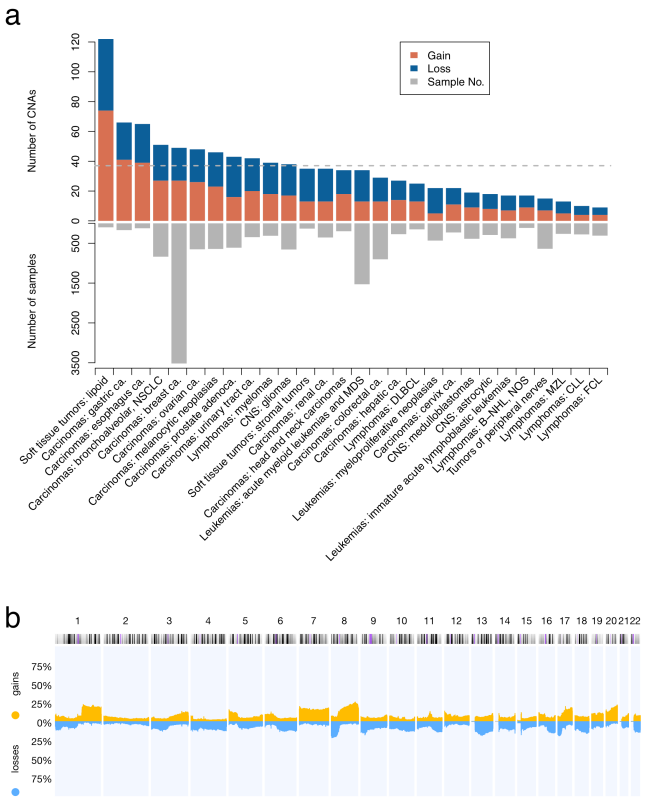


Figure 1. The distribution of somatic CNAs in cancer genomes. (a) Bar graph of input sample number (grey) and average number of genomic gains (red) and losses (blue) per case. Each cancer type is represented by more than 100 samples. The dashed line indicates the average number of the entire dataset. (b) Somatic CNA densities over 1-Mb non-overlapping windows across each of the 22 autosomes calculated by 16,264 cancer samples.

The rearrangement structure of cancer genome based on the total of 606,318 CNAs revealed a number of hotspots (**Figure 1b**). Although these hotspots probably point towards oncogenes or tumor suppressor genes, most CNAs span hundreds of kilo-bases containing multiple genes. Thus, a potential correlation may exist between CNA regions and the fundamental genome structure, which is important for tumorigenesis or the evolution of cancer genome.

Focal CNAs are enriched in gene-rich regions

In order to investigate the possible non-neutral selection for CNAs in cancer, we first divided the genome into 10 Mb non-overlapping intervals and counted the number of genes within each interval. We found that about 10% of the 248 genomic intervals contain twice as many as the average 73 genes in 10 Mb intervals. These intervals can be considered as gene-rich regions. To assess whether CNA hotspots are associated with gene-rich regions, we determined the number of samples that exhibit CNAs in each interval, and considered genomic gains and losses respectively. Since many CNAs encompass thousands of gene loci, far beyond the size of interval, we confined ourselves to focal CNAs, which were defined here as being smaller than 5 Mb. We found that focal genomic gains were significantly positively correlated with gene-rich regions (R = 0.562; Spearman correlation). To evaluate whether the observed correlation could be obtained by chance, we randomly shuffled the number of genes in intervals across the genome, and calculated the correlation coefficient. We repeated this permutation 10,000 times and the P-value < 0.0001. We also observed a positive correlation between focal genomic losses and gene-rich regions (R = 0.342; P-value < 0.0001; Spearman correlation).

We subsequently determined the correlation at different interval sizes (5 Mb and 15 Mb) and CNA sizes (smaller than 2 Mb and 10 Mb). We followed the method described above - i.e., we identified 5 Mb (or 15 Mb) non-overlapping intervals tiling the genome. For each such interval, we counted the number of genes and samples present focal CNAs (smaller than 2 Mb or 10 Mb). As above, we investigated genomic gains and losses respectively. For each combination of these parameters, we found that focal CNAs were significantly enriched in gene-rich regions (**Figure 2**). In general, genomic losses exhibited slightly reduced correlation coefficients compared with genomic gains.

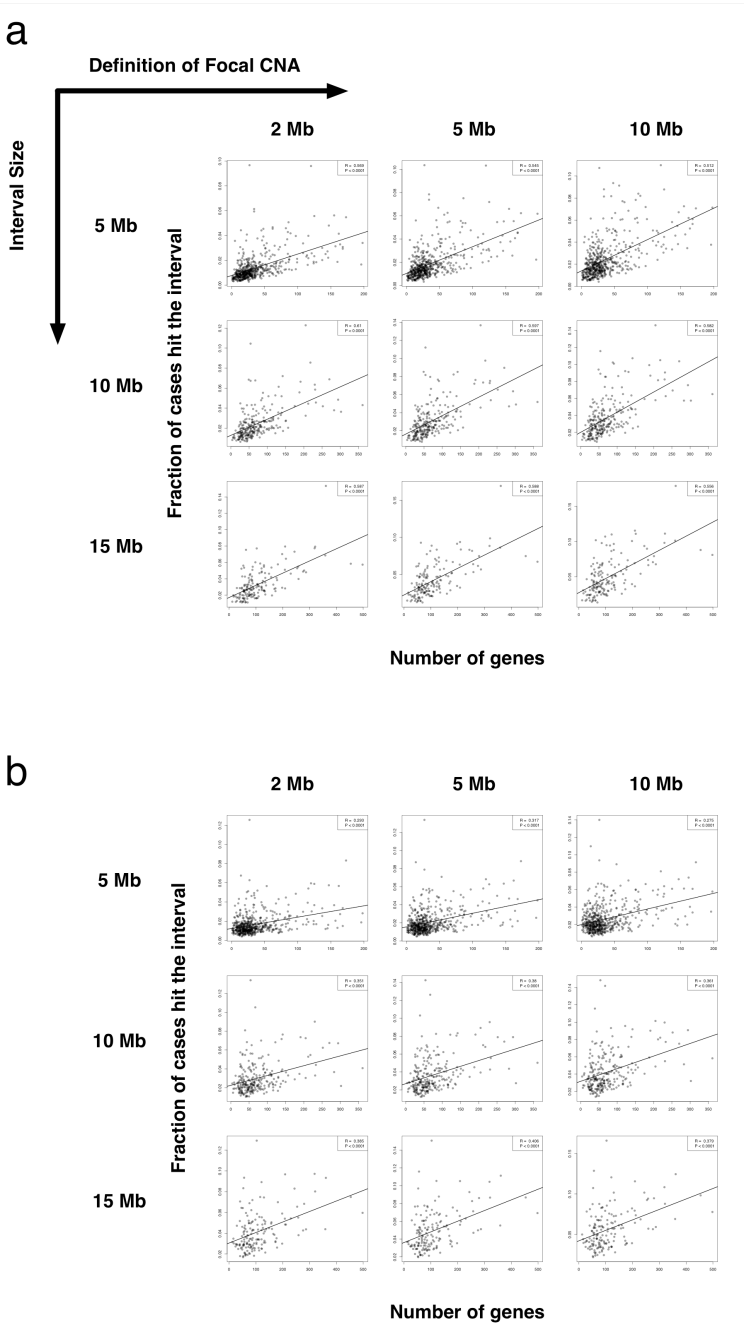


Figure 2. The correlation coefficients between gene density and focal CNA frequency across the genome. (a) Genomic gains. Each point represents a genomic interval. The combination of two parameters, interval size and definition of focal CNA, was considered for calculating Spearman correlation coefficient. The *P*-value is based on 10,000 times permutation. (b) Genomic losses.

Furthermore, genomic regions with high frequency of chromosomal aberrations are likely to contain “driver” mutations, also known as cancer genes. The sporadic CNA regions are usually called “passengers”, which represent random somatic events. To explore whether this positive correlation was only driven by cancer genes, we downloaded the catalogue of cancer gene from COSMIC database, which so far contains 448 annotated cancer genes. The positive association remained even after we excluded all intervals that contain cancer

genes (**Supplementary Figure 1**). These findings suggest that the correlation between CNAs and gene-rich regions is probably genuine.

Since genes are in various sizes, one might expect that coding regions of each interval may better represent functionally important areas. We then performed our analysis based on coding region of intervals instead of gene numbers. The sum of genomic lengths of genes in each interval was calculated. Indeed, similar positive correlations were found between CNA occurrence and coding regions (**Supplementary Figure 2**). Although the correlation was slightly weaker than the condition of gene numbers, it provided similar results from a different angle.

CNA breakpoints are enriched in gene-rich regions

We obtained 1,212,636 CNA breakpoints from the 16,264 cancer genomes. In order to identify the genome-wide distribution of breakpoints, we again divided the genome into 1 Mb non-overlapping intervals and counted the number of breakpoints within each interval. The breakpoints, like CNAs, were also organized in hotspots across the genome (**Figure 3a**). We zoomed in two regions with high and low breakpoint frequencies, respectively, and overlaid genes and their corresponding genomic locations (**Figure 3b**). We found that, of these two regions, the breakpoint hotspot seemed to harbor more genes than the other breakpoint-poor region. This observation led us to investigate the correlation between CNA breakpoints and gene density on a genome-wide scale. We found that breakpoints were positively correlated with gene-rich regions. Furthermore, our observation was independent of the size of intervals. As most CNAs are much larger than 1 Mb, we performed the analysis with 3 Mb, 5Mb, 10Mb and 15Mb intervals (**Figure 3c**).

The correlation of CNAs, breakpoints and gene-rich regions across the whole genome was graphically exhibited by a Circos map (**Figure 4**). The profile of focal CNAs that smaller than 5 Mb and histogram of breakpoints in 3 Mb intervals were plotted. Clusters of CNAs, breakpoints and genes could be clearly observed in specific regions. These data indicated that the highly variable genomic regions were largely correlated with gene density across the entire genome.

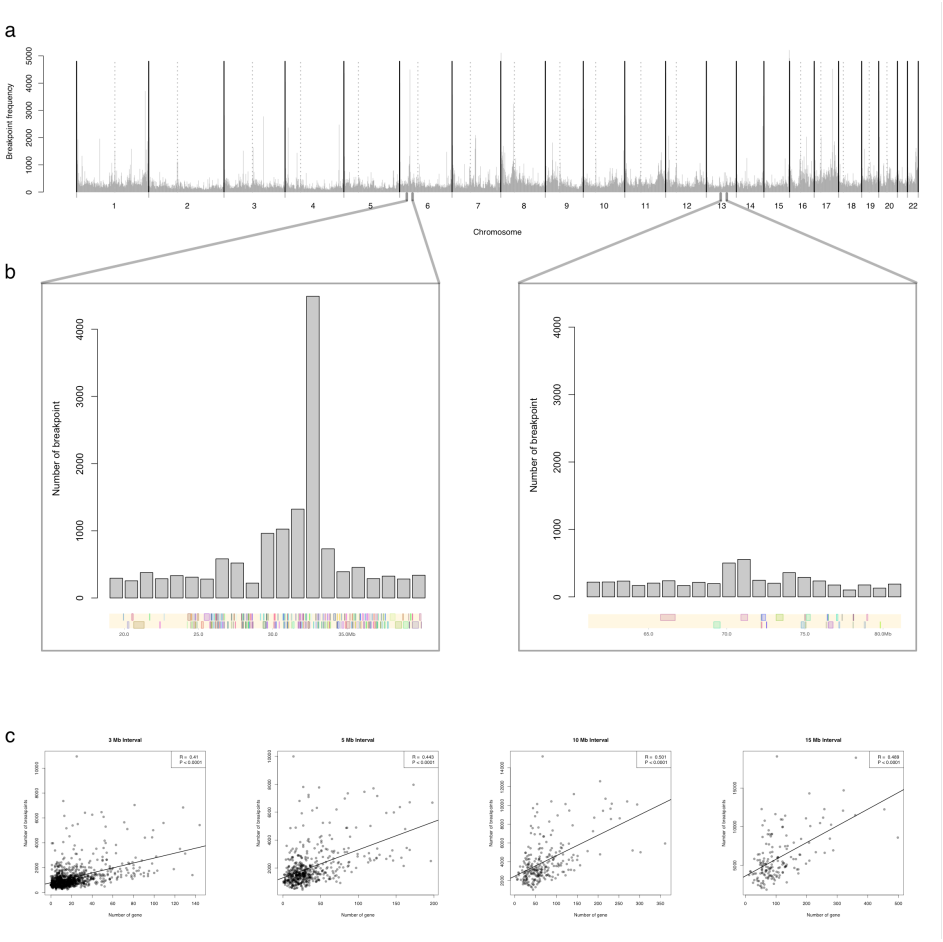


Figure 3. The distribution of DNA breakpoints and its correlation with gene density in cancer genomes. (a) DNA breakpoint densities calculated over 1-Mb non-overlapping windows across the 22 autosomes. Dashed lines indicate centromeres. (b) Two zoomed-in genomic regions represent that high and low breakpoint frequency region overlay gene-rich and gene-poor region, respectively. The colored boxes on x-axis illustrate genes with respect to their sizes and relative positions in the genome. (c) Spearman correlation coefficients between breakpoints and gene density at different genomic intervals.

Arm-level CNAs are negatively correlated with the number of genes covered

Copy number aberrations are frequently classified into “focal” and “arm-level” according to their lengths. Compared to focal alterations, arm-level CNAs are large events involving an entire chromosome arm and cover hundreds of genes. These two types of aberrations might arise through different mechanisms during cancer development. Arm-level CNAs are constantly filtered out when identifying “driver” mutations from CNA hotspots. However, the occurrence and number of whole chromosome arm alteration events are important in specific cancer subtype classification, and may reveal clinical implication. In our study, arm-level CNAs were defined as those with an extension longer than 80% of the respective chromosome arm. Consistent with previous reports, the frequency and distribution of arm-level CNAs identified in our data set were negatively correlated with the

number of genes on related arm (**Figure 5**). In addition, we investigated tumor type specificity of arm-level CNAs. We explored data from 6 tumor types that each has more than 650 samples. In general, the trends of negative correlation were observed in all studied cancer types (**Supplementary Figure 3**). Although the biological significance of this trend is unclear, one possible explanation is that these large events disturb more genes and undergo additional negative selective pressure.

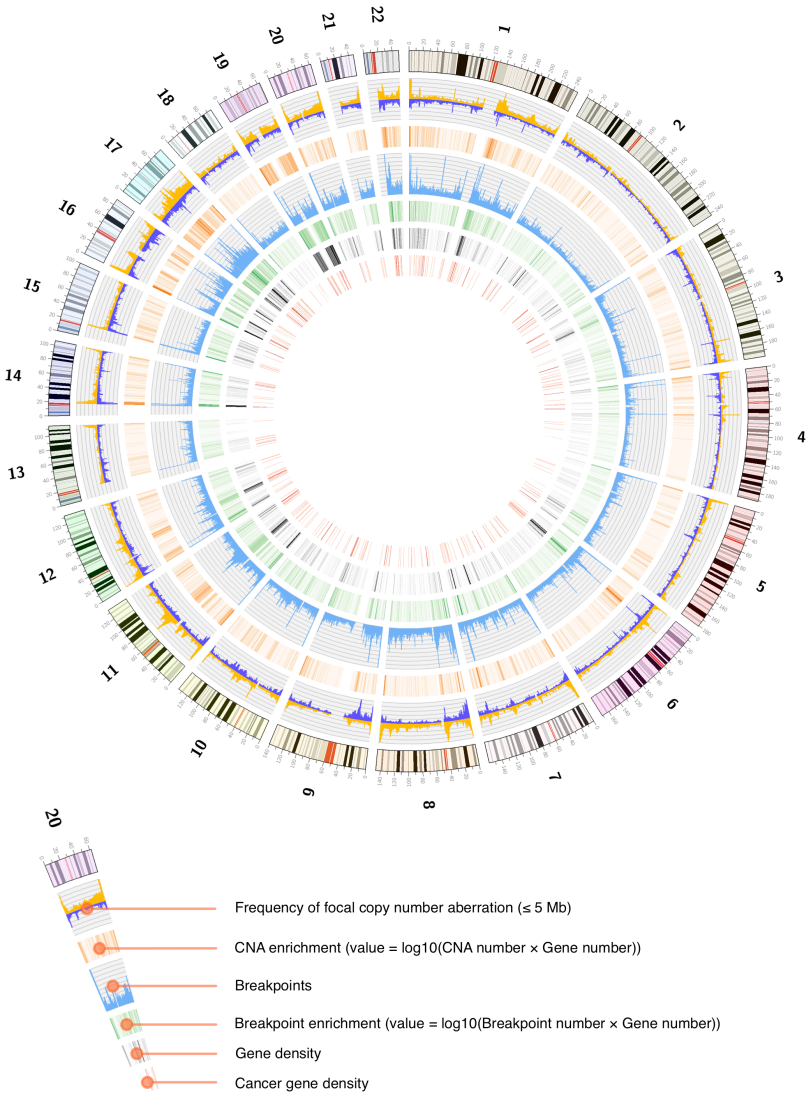


Figure 4. Circular representation of gene density and genomic instability region across the whole genome. Chromosome ideograms are oriented from p-arm to q-arm in a clockwise direction, while centromeres are shown in red (outer circle). The frequency of focal CNA that are smaller than 5 Mb is plotted with gains in yellow and losses in blue (second circle; ranges from 0 to 900 samples). The enrichment between focal CNA and gene density is shown in gradually changing orange per 1 Mb interval (third circle). The histogram of DNA breakpoints is depicted in non-overlapping 1 Mb intervals (fourth circle; ranges from 0 to 2000). The enrichment between breakpoints and gene density is shown in gradually changing green per 1 Mb interval (fifth circle). Gene density across the genome is shown by gradually changing black per 1 Mb interval (sixth circle). The 448 COSMIC cancer genes are represented by gradually changing red per 1 Mb in the innermost circle. All genomic locations are based on the human reference genome UCSC build hg18.

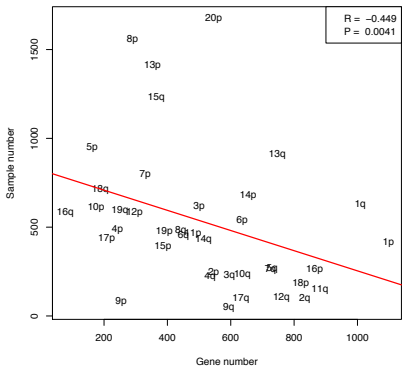


Figure 5. The frequency of arm-level CNAs is negatively correlated with chromosome size. The red line indicates a linear fit to the data. The correlation coefficient is based on Spearman correlation.

Control for robustness of results

In order to control the robustness of our results, we performed several robustness checks. First, centromeres and telomeres are often substrates for rearrangements that are associated with structural genomic alterations in cancer. Therefore, we excluded 1 Mb genomic intervals that overlap with centromeres and telomeres of each chromosome. We then repeated our analysis to determine whether focal CNAs and gene-rich regions were still positively correlated. We found that even after excluding the centromeric and telomeric intervals, both genomic gains and losses were significantly enriched in the gene-rich regions (**Figure 6a,b**). We performed the same method on coding region of intervals, and found that the remaining intervals also exhibited positive correlations (data not shown). Therefore, our observations are not biased by including centromeric and telomeric regions.

Moreover, centromeres and telomeres are also common outliers of CNA breakpoints. For breakpoint analysis, we again excluded all 1 Mb intervals that extended to centromeres and telomeres of each chromosome. As expected, CNA breakpoints were enriched in gene-rich regions at all the 4 analyzed interval sizes (data not shown).

Since our data set is composed of genomic profiles from 180 array platforms, the definitions of focal CNAs according to segment sizes may have potential platform biases. The resolution of a platform depends on the density of probes on the array. To obtain the distribution of segment size in different platforms, we divided these platforms into 3 groups according to their probe numbers and techniques (BAC/P1 and DNA/cDNA, oligonucleotide ≧ 200K and oligonucleotide < 200K). **Figure 6c** shows the density plot of

segment size in each platform group, as well as our definitions of focal CNAs in above analysis. Among the 3 groups, “oligonucleotide ≧ 200K” showed relatively smaller size of alterations compared to other groups. This may be due to the higher sensitivity with increased probe numbers of these platforms. We repeated the correlation analysis on each of these platform groups, and found the same trend: the focal CNAs were significantly enriched in gene-rich regions (**Supplementary Table 3**). Therefore, our conclusions are not biased by the resolution of platforms.

As our somatic CNAs were derived from various cancer types, we then performed cancer type specific analysis. We focused on 6 clinical groups that each has more than 650 samples in our data set. All these cancer types indicated a similar positive correlation between focal CNAs and gene-rich region (**Supplementary Table 4**). In general, genomic losses showed weaker correlations compared to genomic gains. Moreover, the significance of correlations were distinct in different cancer types. It may caused by the inherit features of specific cancers, such as segment size and complexity of genome-wide copy number profiles. Taken together, our data indicate that genomic instability regions are enriched in gene-rich regions across the whole genome.

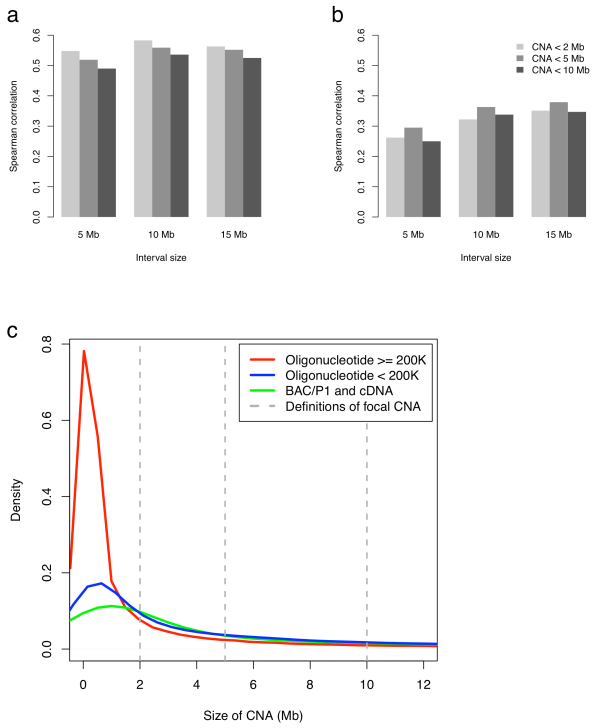


Figure 6. Robustness checks of results. a-b, Correlations after removing telomere and centromere regions calculated by genomic gains (a) and losses (b). (c) Density plot of CNA size across 3 different platform groups. The dashed lines represent definitions of focal CNA that are used in the main text (2 Mb, 5 Mb and 10 Mb).

Discussion

Here we have presented that both genes and CNAs are often clustered into hotspots, and have explored the underlying correlation between gene distribution and copy number alteration profiles across cancer genomes. To achieve this, we collected more than 16,000 cancer samples from 3 public resources of microarray data sets. Notably, focal CNAs were significantly enriched in gene-rich regions. As another manifestation of genome instability, DNA breakpoints also followed this trend. It provided us a global insight into the relationship between cancer genome instability and structure from a new perspective. The enrichment revealed that there is a non-neutral selection pressure for CNA regions across the genome. Due to observed heterogeneity of CNAs in cancer genomes, individual tumor probably follow a distinct path towards tumorigenesis. Many genes in CNAs may lose their functions or take on new roles to promote clone expansion. Genome instability could generate enough variation on which these non-neutral selection events can operate during tumor evolution. A full understanding of how these events contribute to specific tumor will require further studies to investigate the differential expression of genes in CNA regions.

A negative correlation between arm-level CNAs and the size of arms was observed from the entire data set, consistent with previous studies. To further confirm the negative correlation with tumor specific data, we looked into arm-level CNAs of 6 cancer types. The underlying mechanism for this observation is unknown. It may reveal additional negative selective pressure on gene-rich arms, or inherent low arm-level CNA ratios of these arms.

To avoid bias, we performed cancer type and platform specific analysis across the entire dataset. Generally, focal CNAs in most cancers presented enrichment in gene-rich regions, although the extent of enrichment was a little bit different. A couple of cancers, such as hematological malignancies, presented an inherent low-level of copy number changes, especially of focal CNAs. Accordingly, these cancers showed relatively low correlation coefficients. Moreover, our data set was generated from 180 array platforms with various resolutions. Due to increased probe numbers, high resolution platforms are more sensitive to small copy number changes, thus usually show a substantially higher number of alterations. Therefore, compared to low resolution platforms, a higher proportion of segments that derived from high resolution arrays fitted our definitions of focal CNAs. For example, in the definition of 5 Mb, we got ~52% segments of BAC/P1 and DNA/cDNA

arrays, while this number increased to ~80% in oligonucleotide arrays that contain more than 200K probes. Nevertheless, this did not affect the consequence.

In conclusion, through large-scale oncogenomic array data analysis, our results revealed a significant positive correlation between genomic instability regions and gene distribution in cancer genomes. It may enable a better elucidation of mechanisms by which CNAs contribute to tumor development, and eventually promote a more systematic understanding of human cancer.

Methods

Data sets of cancer genome

We collected annotated human cancer genome data from arrayMap database, including the normalized probe intensities, segmented data and quality information. These 19,471 arrays were generated from 390 studies from 3 publicly available data sources, including NCBI GEO, EBI ArrayExpress and supplemental materials of publications. In this study, all collected arrays were high quality genome-wide human cancer samples. In the case of technical repeats, such as different platforms for one sample, only one of the arrays was considered for analysis (preferably with the highest resolution and/or best overall quality). After this filtering process, 16,264 samples were left in total. These data were generated by 180 platforms with various techniques, resolutions and probe lengths. Detailed information for the input samples and classification of platforms are shown in **Supplementary Table 2**.

Data processing

For Affymetrix arrays, the aroma.affymetrix R package was employed to generate log2 scale probe level data from original CEL files. For non-Affymetrix arrays, available probe intensity files were downloaded from GEO or ArrayExpress and processed. For supplemental materials of publications, all the probe-level log2 ratio profiles or segmented data were converted to an unique format and stored into database. Segmentation was performed by the CBS (Circular Binary Segmentation) algorithm. The thresholds for calling genomic gains and losses were obtained from publication of individual study, or if it is not available, empirically assigned. In our study, the boundaries of CNAs were defined as breakpoints. The probe locations were mapped on the human reference genome (UCSC build hg18). The processed array profiles can be visualized and downloaded through the arrayMap website (www.arraymap.org).

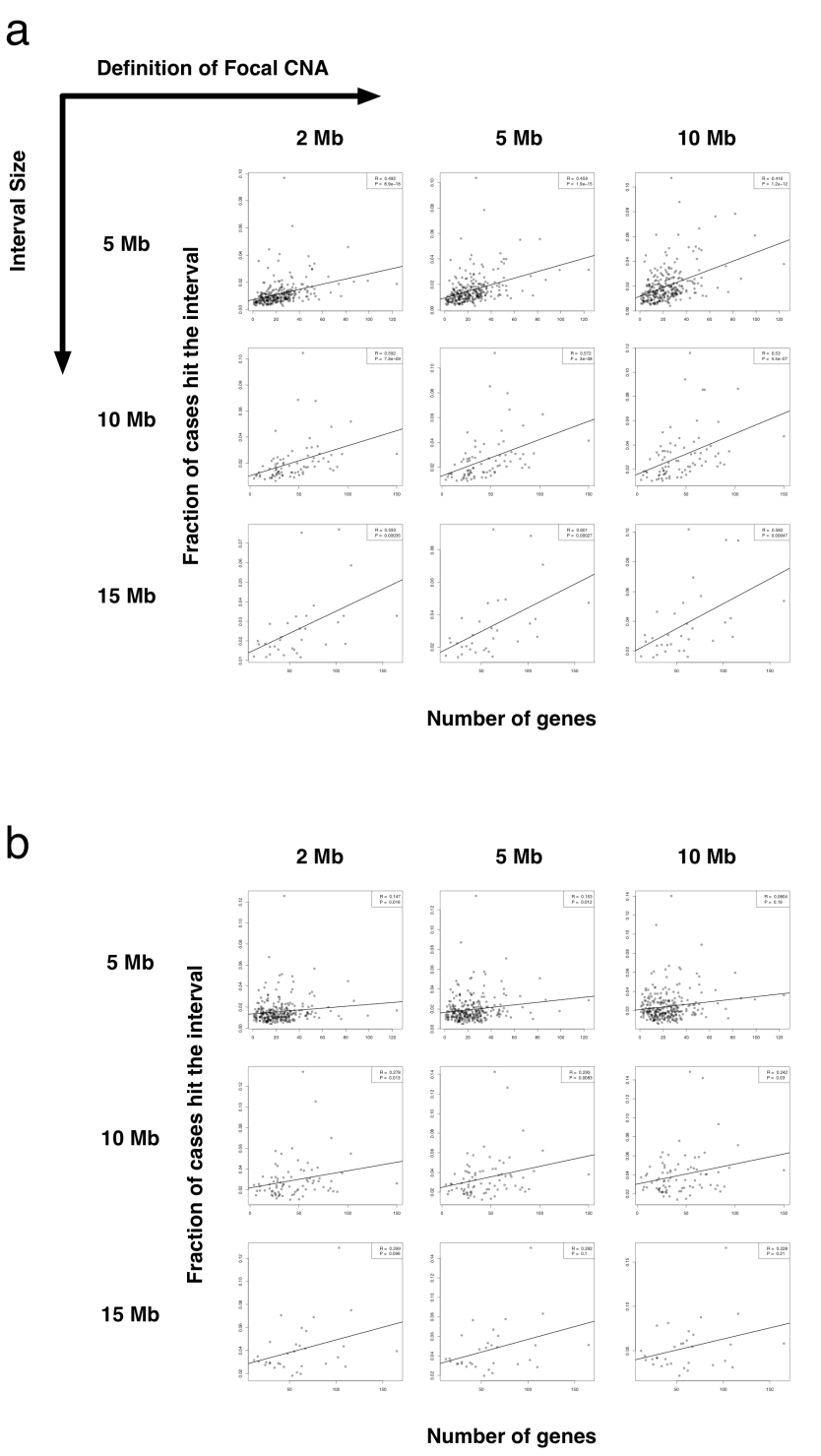
Ensembl gene and cancer gene

We downloaded the Ensembl gene list from Biomart (release 54). This gene list was processed to obtain genes with unique combination of Ensembl gene identifier (Ensembl

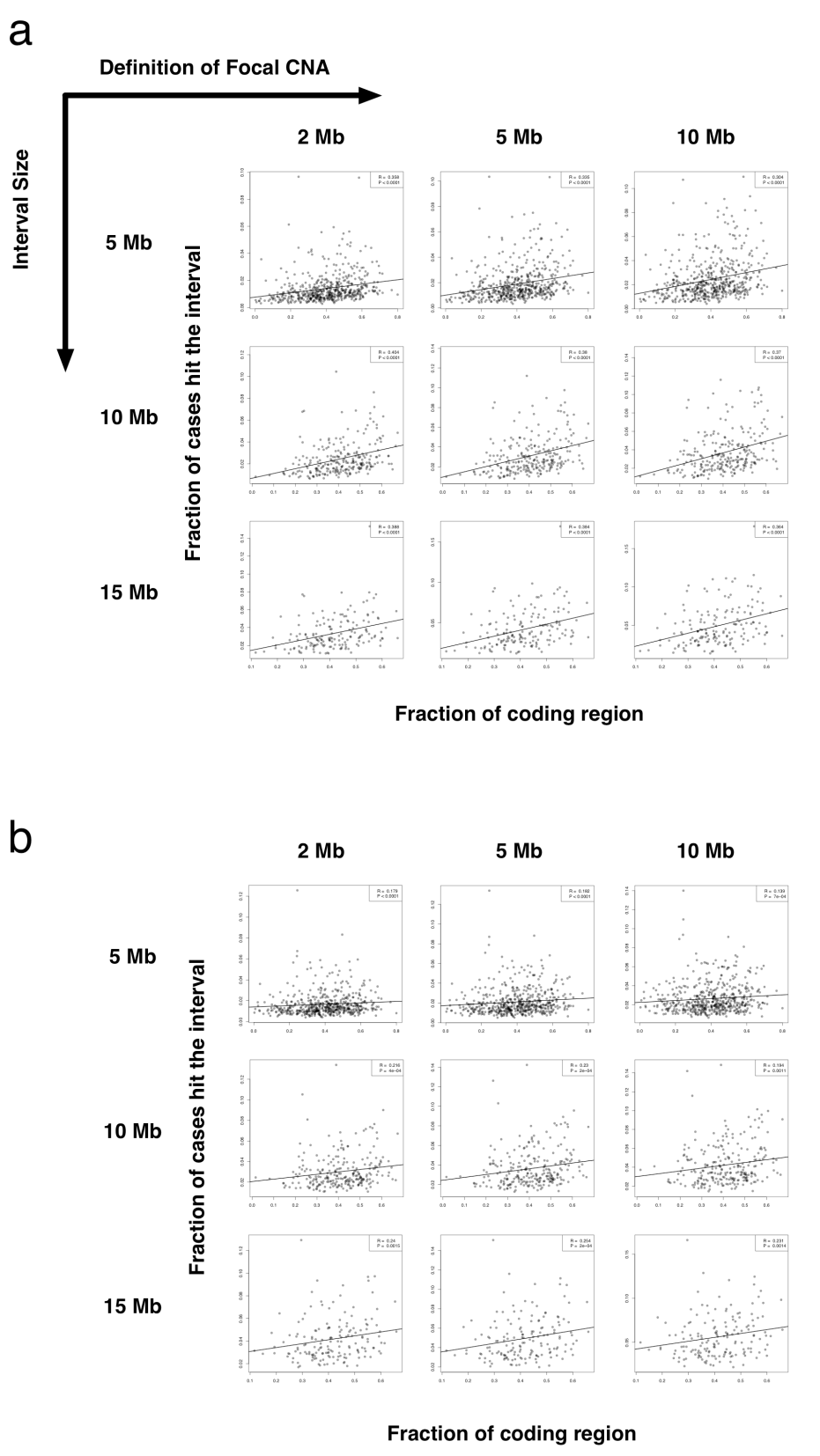
id) and start/end positions in the genome. Mitochondrial and sex chromosomal genes were removed, and resulting in a total of 20,207 genes (**Supplementary Table 5**).

The catalogue of cancer gene was obtained from the COSMIC database (v61 release). The gene positions were converted from gene symbols by UCSC Genome Browser (397 genes) or from Entrez gene IDs by NCBI Gene (51 genes). In total, we got 448 cancer genes (**Supplementary Table 6**). All mapped genomic locations of Ensembl and COSMIC cancer genes were based on UCSC build hg18.

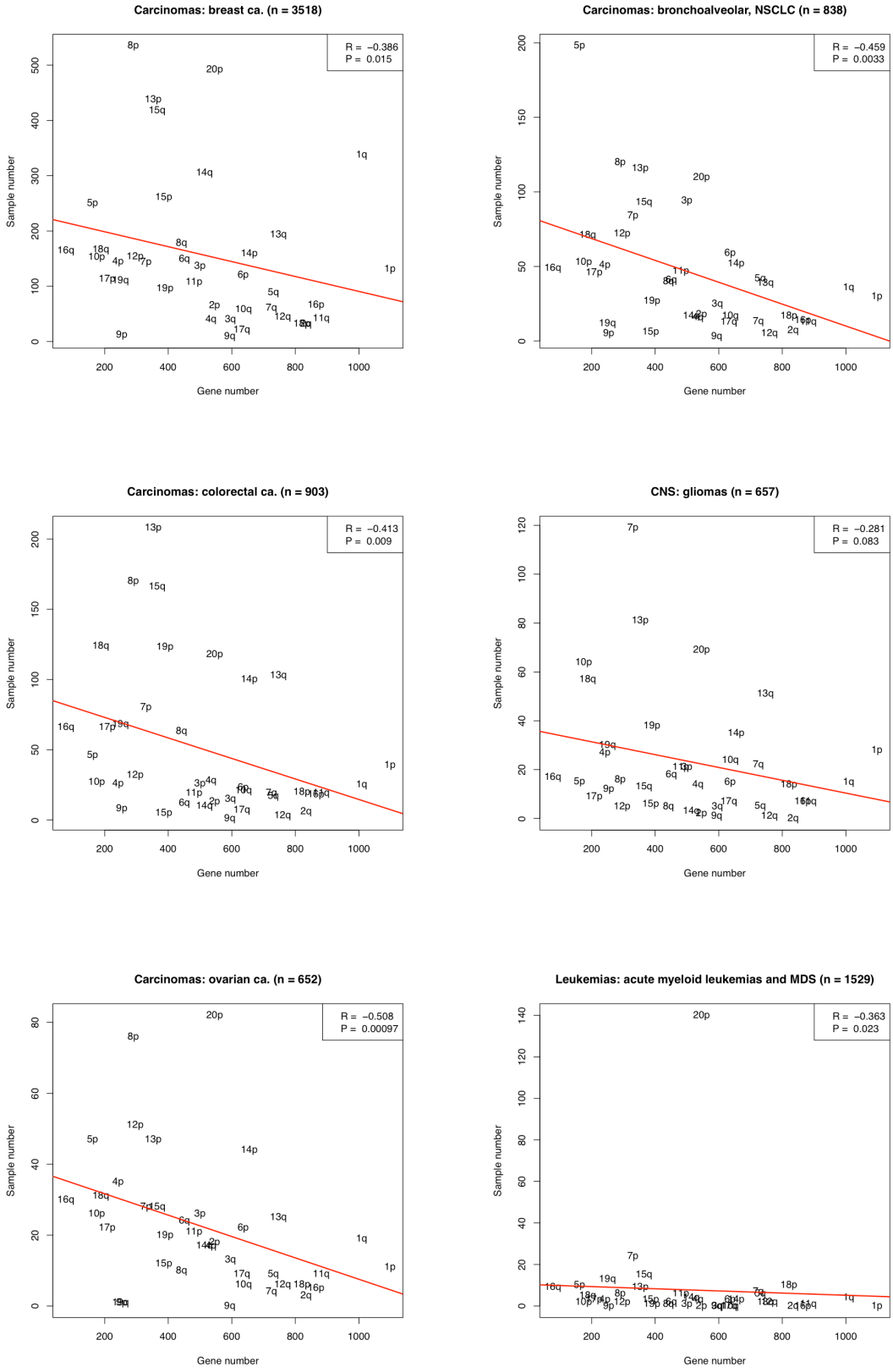
Supplementary Figures



Supplementary Figure 1. The correlation coefficients between gene density and focal CNA frequency among intervals without cancer gene. Each point represents a genomic interval that without COSMIC cancer gene. Genomic gains (a) and losses (b) were considered separately. The coefficient was calculated by Spearman correlation. The *P*-value is based on 10,000 times permutation.



Supplementary Figure 2. The correlation coefficients between gene density and focal CNA frequency in terms of coding region. Each point represents a genomic interval. For each interval, the fraction of coding region was obtained. Genomic gains (a) and losses (b) were considered separately. The coefficient was calculated by Spearman correlation. The *P*-value is based on 10,000 times permutation.



Supplementary Figure 3. The negative correlation between frequency of arm-level CNAs and chromosome size among different cancer types. Each cancer type is represented by more than 650 samples. Red lines indicate linear fits to the data. The correlation coefficient is based on Spearman correlation.

Supplementary Tables

Supplementary Table 1. Overview of input data set

Category	Array-level	Case-level
Total number	19471	16264
Series	390	390
Platform	184	180
Cancer type (ICD-O)	135	135
Cancer type (diagnostic group)	62	62
Source (primary)	17043	14340
Source (cell line)	2085	1604

Supplementary Table 3. Platform type specific enrichment

		Gain						Loss					
	Focal CNA size	2 Mb		5 Mb		10 Mb		2 Mb		5 Mb		10 Mb	
Platform type	Interval size	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
BAC/P1 and DNA/cDNA	5 Mb	0.63	< 0.0001	0.614	< 0.0001	0.566	< 0.0001	0.497	< 0.0001	0.497	< 0.0001	0.45	< 0.0001
	10 Mb	0.689	< 0.0001	0.688	< 0.0001	0.659	< 0.0001	0.551	< 0.0001	0.56	< 0.0001	0.533	< 0.0001
	15 Mb	0.693	< 0.0001	0.681	< 0.0001	0.657	< 0.0001	0.594	< 0.0001	0.592	< 0.0001	0.566	< 0.0001
Oligonucleotide < 200K	5 Mb	0.35	< 0.0001	0.351	< 0.0001	0.372	< 0.0001	0.2	< 0.0001	0.197	< 0.0001	0.108	0.0061
	10 Mb	0.402	< 0.0001	0.399	< 0.0001	0.417	< 0.0001	0.222	0.0001	0.252	0.0001	0.19	0.0009
	15 Mb	0.386	0.0001	0.377	< 0.0001	0.351	< 0.0001	0.285	0.0003	0.29	0.0001	0.192	0.0079
Oligonucleotide ≥ 200K	5 Mb	0.536	< 0.0001	0.521	< 0.0001	0.498	< 0.0001	0.21	< 0.0001	0.248	< 0.0001	0.238	< 0.0001
	10 Mb	0.583	< 0.0001	0.581	< 0.0001	0.572	< 0.0001	0.276	< 0.0001	0.303	< 0.0001	0.302	< 0.0001
	15 Mb	0.561	< 0.0001	0.56	< 0.0001	0.551	< 0.0001	0.294	< 0.0001	0.308	< 0.0001	0.301	< 0.0001

Supplementary Table 4. Cancer type specific enrichment

		Gain						Loss					
	Focal CNA size	2 Mb		5 Mb		10 Mb		2 Mb		5 Mb		10 Mb	
Cancer type	Interval size	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
Breast ca.	5 Mb	0.563	< 0.0001	0.534	< 0.0001	0.495	< 0.0001	0.341	< 0.0001	0.369	< 0.0001	0.376	< 0.0001
	10 Mb	0.608	< 0.0001	0.592	< 0.0001	0.575	< 0.0001	0.399	< 0.0001	0.438	< 0.0001	0.453	< 0.0001
	15 Mb	0.616	< 0.0001	0.599	< 0.0001	0.573	< 0.0001	0.474	< 0.0001	0.501	< 0.0001	0.504	< 0.0001
Colorectal ca.	5 Mb	0.261	< 0.0001	0.324	< 0.0001	0.376	< 0.0001	0.181	< 0.0001	0.281	< 0.0001	0.309	< 0.0001
	10 Mb	0.367	< 0.0001	0.434	< 0.0001	0.48	< 0.0001	0.257	< 0.0001	0.35	< 0.0001	0.367	< 0.0001
	15 Mb	0.393	< 0.0001	0.427	< 0.0001	0.47	< 0.0001	0.275	0.0005	0.358	< 0.0001	0.372	< 0.0001
AML and MDS	5 Mb	0.299	< 0.0001	0.324	< 0.0001	0.354	< 0.0001	0.057	0.102	0.002	0.477	-0.114	0.995
	10 Mb	0.289	< 0.0001	0.311	< 0.0001	0.34	< 0.0001	0.072	0.072	0.042	0.25	-0.053	0.803
	15 Mb	0.251	0.0001	0.236	0.0016	0.191	0.0084	0.086	0.138	0.059	0.228	-0.073	0.824
NSCLC	5 Mb	0.244	< 0.0001	0.249	< 0.0001	0.25	< 0.0001	0.163	< 0.0001	0.266	< 0.0001	0.317	< 0.0001
	10 Mb	0.364	< 0.0001	0.38	< 0.0001	0.382	< 0.0001	0.306	< 0.0001	0.378	< 0.0001	0.43	< 0.0001
	15 Mb	0.398	< 0.0001	0.398	< 0.0001	0.38	< 0.0001	0.277	< 0.0001	0.347	< 0.0001	0.37	< 0.0001
Gliomas	5 Mb	0.541	< 0.0001	0.516	< 0.0001	0.486	< 0.0001	0.337	< 0.0001	0.378	< 0.0001	0.376	< 0.0001
	10 Mb	0.646	< 0.0001	0.636	< 0.0001	0.611	< 0.0001	0.341	< 0.0001	0.387	< 0.0001	0.411	< 0.0001
	15 Mb	0.639	< 0.0001	0.623	< 0.0001	0.603	< 0.0001	0.381	< 0.0001	0.427	< 0.0001	0.417	< 0.0001
Ovarian ca.	5 Mb	0.467	< 0.0001	0.464	< 0.0001	0.449	< 0.0001	0.177	< 0.0001	0.17	< 0.0001	0.118	0.0028
	10 Mb	0.517	< 0.0001	0.501	< 0.0001	0.497	< 0.0001	0.208	0.0003	0.216	0.0006	0.182	0.0029
	15 Mb	0.481	< 0.0001	0.488	< 0.0001	0.48	< 0.0001	0.29	< 0.0001	0.274	0.0001	0.249	0.0009

AML, acute myeloid leukemia; MDS, myelodysplastic syndrome; NSCLC, Non-small-cell lung carcinoma

6 CHROMOTHRIPSIS: CHROMOSOME CATASTROPHES

6.1 PREFACE

Somatically acquired genomic rearrangements may result in complex patterns of regional copy number changes, and consequently contribute to cancer development. Analysis of genomic rearrangements will promote understanding of the biological mechanisms of oncogenesis[2,3]. Through the decades efforts of the cancer research community, the stepwise development of cancer with the gradual accumulation of multiple genetic alterations has been the most widely accepted model[3]. In this model, somatic mutations are classified into two categories, “driver” mutations which contribute to the tumor’s progression, and “passenger” mutations which do not have any effect on the clonal expansion. Both types of mutations may be acquired during normal cell division, reflecting the intrinsic mutations. During the process of cancer development, natural selection acts on the phenotypic diversity and may weed out cells carrying deleterious mutations, or foster cells that have acquired selective advantage. Within a cancer genome, there are probably hundreds of somatic mutations. The procedure may take years to convert a normal cell to a tumor cell.

Recently, using state-of-the-art genome analysis techniques, a phenomenon termed “Chromothripsis” was characterized in cancer genomes, defined by the occurrence of tens of hundreds of clustered genomic rearrangements, supposedly having arisen in a single catastrophic event[105]. In this event, contiguous chromosomal regions are fragmented into many pieces via presently unknown mechanisms[105,106]. Supposedly, the cell’s DNA repair machinery randomly fuses these segments together to rescue the genome. Since this is a one-off event, it has been proposed that this procedure could provide an alternative oncogenetic route, in contrast to the step-by-step paradigm of cancer development.

Furthermore, chromothripsis may be a widespread mechanism in cancer progression, as recent researches have reported chromothripsis events in germline and non-human genomes[118,119]. This interesting observation has stimulated many following studies after

the initial report. Subsequently, chromothripsis was found in breast cancer, colorectal cancer, leukaemia, medulloblastoma, prostate cancer, and so on[121-132]. However, due to the relatively low incidence of this phenomenon, most studies are limited to relatively small numbers of observed events. Larger sample numbers are required to gain further insights into features and mechanisms of these events in different cancers.

6.2 CHARACTERIZATION OF CHROMOTHRIPSIS

In this work, I designed a statistical model to automatically detect chromothripsis-like patterns from genomic array data. The input dataset is more than 22000 high quality arrays from arrayMap database. The previous studies have proven that the identification of chromothripsis patterns from genomic array data is feasible. The algorithm is based on scan statistics and two related features - copy number status change times and breakpoints clustering[195]. The pipeline of this project is shown in Figure 8. Applying the methodology to the input dataset, I detected 918 chromothripsis-like cases. With these data, I was able to determine the frequency and genomic distribution of chromothripsis events.

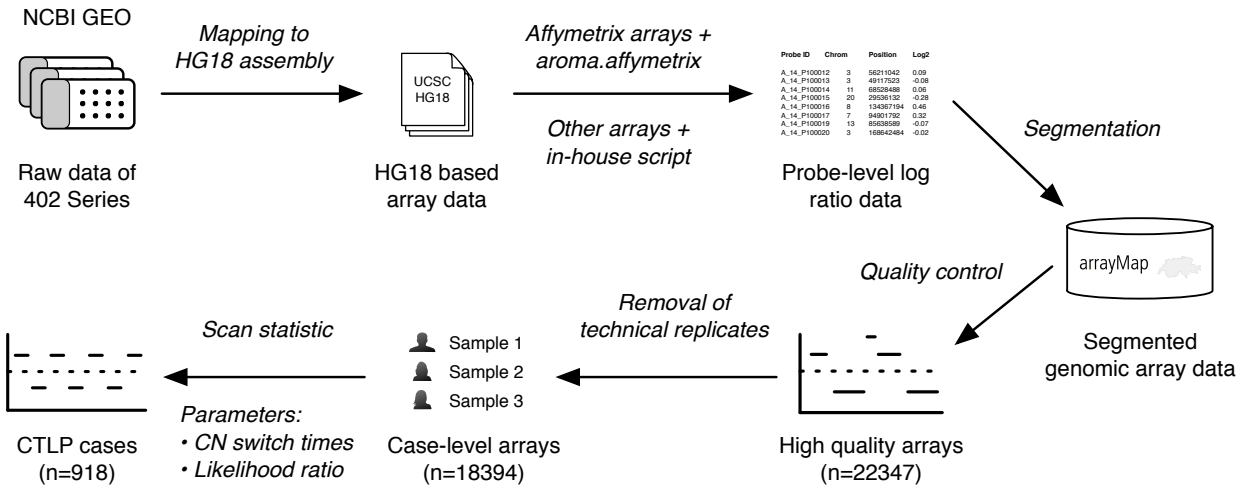


Figure 8. Schematic description of the chromothripsis-like patterns detection procedure. Raw array data of 402 GEO series are first collected and re-analyzed, then annotated and stored in arrayMap database. For high quality arrays, a scan-statistic based algorithm was employed to identify CTLP cases.

First, I found that chromothripsis events exhibited an uneven distribution along the tumor genomes, with disease related local enrichment. Fragmentation hotspots were found to be located on chromosome 8, 11, 12 and 17. These chromosomal pulverization regions may reveal associations between tumor type related cancer associated genes and molecular mechanisms behind chromothripsis events. This potential correlation is exemplified by a recently published study that showed the prevalence of mutant *TP53* in chromothriptic Li-Fraumeni syndrome associated Sonic-Hedgehog medulloblastomas. Then, as the collection of array data represents 132 cancer types, the incidences of chromothripsis in diverse tumor types were able to be estimated. Interestingly, soft-tissue tumors exhibited particularly high chromothripsis frequencies. This finding supports and improves upon a previous prediction of particularly high chromothripsis rate in bone tumors. In addition, genomic context analysis revealed that chromothripsis rearrangements frequently occurred in genomes that additional harbored multiple copy number aberrations. It is possible to contradict the proposed singular “shortcut” to cancer genome generation. Plausible and non-exclusive explanations could be that chromothripsis might frequently arise due to previously established errors in the maintenance of genomic stability, or that chromothriptic aberrations involving genomic maintenance genes may predispose to the acquisition of additional CNA. For those frequent cases exhibiting additional non-chromothripsis CNA events, their possible contribution to oncogenesis has to be considered when modeling the role of chromothripsis in cancer development. Moreover, an investigation into the affected chromosomal regions showed a large proportion of arm-level pulverization and telomere related events, which would support breakage-fusion-bridge cycles as one of the potential underlying mechanisms. Finally, I evaluated clinical associations of chromothripsis, based on the clinical information at hand. Chromothripsis seemed to occur at a more advanced patient age as compared to non-chromothripsis samples. It mainly occurred at stage II and III, which was significantly different from the stage distribution of total samples. As for tumor grade, chromothripsis showed a predominance for grades 2 and 3. I also found that this phenomenon was overrepresented in cell lines compared to primary tumors. According to follow-up information, patients with chromothripsis survived a significantly shorter time than those without this phenomenon. Overall, this project characterized heterogeneous features of chromothripsis through a large-scale analysis of oncogenomic arrays data sets and provides a better understanding of this new paradigm in cancer development.

The manuscript is included below.

Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genomes

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C. Bagheri[3], Christian von Mering[1,2], Mark D. Robinson[1,2], and Michael Baudis[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

[2]Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

[3]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

ABSTRACT

Chromothripsis is a newly discovered type of genomic rearrangement, characterized by locally clustered copy number aberrations. It has been proposed that it may arise during a single genome shattering event. This could provide an alternative paradigm in cancer development, replacing the gradual accumulation of genomic changes with a ”one-off” catastrophic event. However, the underlying mechanisms and its specific impact on tumorigenesis are still poorly understood. Here, we identified chromothripsis-like genome patterns (CTLP) in 918 cancer samples, from a dataset of more than 22,000 oncogenomic arrays including 132 cancer types. Fragmentation hotspots were found to be located on chromosome 8, 11, 12 and 17. Among the various cancer types, soft tissue tumors exhibited particularly high CTLP frequencies. Genomic context analysis revealed that CTLP rearrangements frequently occurred in genomes harboring multiple additional copy number aberrations (CNAs). An investigation into the affected chromosomal regions showed a large proportion of arm-level pulverization and telomere related events, which would support breakage-fusion-bridge cycles as one of the potential underlying mechanisms. We also report evidence that this catastrophic event may be correlated with patient age, stage and survival rate.

INTRODUCTION

One hallmark of human cancer genomes are somatically acquired genomic rearrangements, which sometimes may result in complex patterns of regional copy number changes (Albertson et al., 2003; Hanahan and Weinberg, 2011). These alterations have the potential to interrupt or activate multiple genes, and consequently have been implicated in cancer development (Yates and Campbell, 2012). Analysis of genomic rearrangements is essential for understanding the biological mechanisms of oncogenesis and to determine rational points of pharmacological interference (Chen et al., 2010; Chin and Gray, 2008; Hanahan and Weinberg, 2011). Some large-scale efforts have been made to correlate genomic rearrangements to genome architecture as well as to the progression dynamics of cancer genomes (Beroukhim et al., 2031; Kim et al., 2012). At the moment, the stepwise development of cancer with the gradual accumulation of multiple genetic alterations is the most widely accepted model (Stratton et al., 2009).

Recently, using state-of-the-art genome analysis techniques, a phenomenon termed ”Chromothripsis” was characterized in cancer genomes, defined by the occurrence of tens to hundreds of clustered genomic rearrangements, supposedly arising in a single catastrophic event (Stephens et al., 2011). In this model, contiguous chromosomal regions are fragmented into many pieces, through presently unknown mechanisms. Supposedly, these segments are then randomly fused together by the cell’s DNA repair machinery. It has been proposed that this ”shattering” and aberrant repair of a multitude of DNA fragments could provide an alternative oncogenetic route, in contrast to the step-by-step paradigm of cancer development (Kitada et al., 2011; Stephens et al., 2011; Stratton et al., 2009). The initial study reported in overall 24 chromothripsis cases, with some evidence of a high prevalence in bone tumors (Stephens et al., 2011).

Besides human cancers, recent studies reported chromothripsis events in germline and non-human genomes (Chiang et al., 2012; Deakin et al., 2012; Kloosterman et al., 2011a). However, due to the overall low incidence of this phenomenon, most studies were limited to relatively small numbers of observed events. For example, in a study screening 746 multiple myelomas by SNP arrays, only 10 cases with chromothripsis were detected (Magrangeas et al., 2011). Larger sample numbers are required to gain further insights into features and mechanisms of these events in different cancers.

The identification of chromothripsis patterns from genomic array data has been considered feasible, according to the techniques applied in previous studies (Kim et al., 2012; Magrangeas et al., 2011; Northcott et al., 2012; Stephens et al., 2011). Table 1 provides an overview of studies which so far have reported instances of chromothripsis in human cancers. Here, we present a statistical model to discover chromothripsis-like (CTLP) from genomic array data sets. Applying our methodology to 22,347 genomic arrays from 402 GEO (Gene Expression Omnibus) derived experimental series (Barrett et al., 2013), we were able to detect 918 chromothripsis-like cases, and to determine the frequency and genomic distribution of CTLP events in this dataset.

Our collection of oncogenomic array data represents 132 cancer types as defined using the ICD-O 3 (International Classification of Diseases for Oncology) coding scheme, enabling us to estimate the incidence of chromothripsis events in diverse tumor types. Among the CTLP cases, varying distributions of fragmented chromosomal regions and large non-CTLP copy number aberrations (CNA) regions were found, and the genomic context of chromothripsis events was investigated. Finally, we evaluated some clinical associations of chromothripsis, based on the clinical information at hand. Overall, this study characterized heterogeneous features of chromothripsis through a large-scale analysis of oncogenomic arrays data sets and provides a better understanding of this new paradigm in cancer development.

RESULTS

Detection of chromothripsis-like patterns from oncogenomic arrays

We collected 402 GEO series, including 22,347 high quality genomic arrays of human cancer samples. The annotated information of arrays, including normalized probe intensity, segmentation data and quality evaluation, was obtained from our arrayMap database (Cai et al., 2012) (see Methods for array processing pipeline). After removing technical repeats (e.g. multiple platforms for one sample), a total of 18,394 cases represent 132 cancer types remained. The input data is summarized, at array and case-level, respectively, in

Supplemental Tables 1 and 2. The segmentation data and array profiling can be accessed and visualized through the arrayMap website (www.arraymap.org).

Table 1 Summary of chromothripsis-like cases identified in previous and current studies

Study[a]	Chromothripsis -like cases[b]	Sample size	Techniques	Cancer/sample types
Stephens et al., 2011	24	776	paired-end sequencing, SNP array	55 cancer types[d]
Kloosterman et al., 2011	1	3[c]	mate-pair sequencing	germline, congenital defects
Le et al., 2011	1	21	aCGH	chordoma
Magrangeas et al., 2011	10	764	SNP array	multiple myeloma
Bass et al., 2011	3	9	whole-genome sequencing	colorectal adenocarcinoma
Kloosterman et al., 2011	4	4	mate-pair sequencing, SNP array	colorectal cancer
Zhang et al., 2012	3	12	whole-genome sequencing	acute lymphoblastic leukaemia
Kitada et al., 2012	5	150	aCGH	na
Poaty et al., 2012	1	14	aCGH	gestational choriocarcinoma
Rausch et al., 2012	52	605	whole-genome sequencing, SNP array	7 cancer types
Jiang et al., 2012	1	4	paired-end sequencing	hepatocellular carcinoma
Molenaar et al., 2012	16	87	paired-end sequencing, SNP array	neuroblastoma
Chiang et al., 2012	2	52	whole-genome sequencing, aCGH	germline
Lapuk et al., 2012	1	6	whole-genome/transcriptome sequencing, aCGH	neuroendocrine prostate cancer
Berger et al., 2012	2	25	whole-genome sequencing	melanoma
Natrajan et al., 2012	1	2	whole-genome sequencing	breast cancer
Nik-Zainal et al., 2012	3	21	whole-genome sequencing	breast cancer
Kloosterman et al., 2012	10	10	mate-pair sequencing, SNP array	congenital disease
Wu et al., 2012	3	3	paired-end sequencing, aCGH	prostate cancer
Northcott et al., 2012	na	1087	SNP array, whole-genome sequencing	medulloblastoma
Jones et al., 2012	2	3	whole-genome sequencing	medulloblastoma
Kroef et al., 2012	1	61	SNP array	multiple myeloma
Govindan et al., 2012	1	17	whole-genome sequencing	non-small cell lung cancer
Kim et al., 2012	124	8227	aCGH, SNP array	30 cancer types
Zehentner et al., 2012	1	28	aCGH	plasma cell neoplasia
Current study	918	18394	aCGH, SNP array	132 cancer types[e]

[a] Data up to 21st December, 2012; [b] na, not available; [c] Family trio: father, mother, child; [d] According to site and histology; [e] Classified by ICD-O code

According to previous studies, segmental copy number status changes and significant breakpoint clustering are two relevant features of chromothripsis (Rausch et al., 2012; Stephens et al., 2011). For an automatic identification of CTLP, we developed a scan-statistic based algorithm (Naus, 1965). We employed a maximum likelihood ratio score, which is commonly used to detect clusters of events in time and/or space and to determine their statistical significance (Kulldorff, 1997) (see Methods). For each chromosome, the algorithm uses a series of sliding windows to identify the genomic region with the highest likelihood ratio as CTLP candidate. In order to test the performance of the algorithm, 23 published chromothripsis cases with available raw array data were collected and used as a

training set. This data contained 31 chromothriptic and 475 non-chromothriptic chromosomes that acted as positive and negative controls, respectively (Supplemental Table 3). Comparison of copy number status change times and likelihood ratios showed that chromothriptic chromosomes could reliably be distinguished from non-chromothriptic ones (Supplemental Fig. 1). We generated a receiver operating characteristic (ROC) curve from the training set results, and selected optimal cutoff values based on this curve (copy number status switch times ≥ 20 and log10 of likelihood ratio ≥ 8) (Fig. 1B). Furthermore, the sliding window scan statistic accurately identified the involved genomic regions (Supplemental Fig. 2). Applying this algorithm to the complete input data set, a total of 1,269 chromosomes from 918 cases passed our thresholds and were marked as CTLP events (Supplemental Fig. 3; Supplemental Table 4).

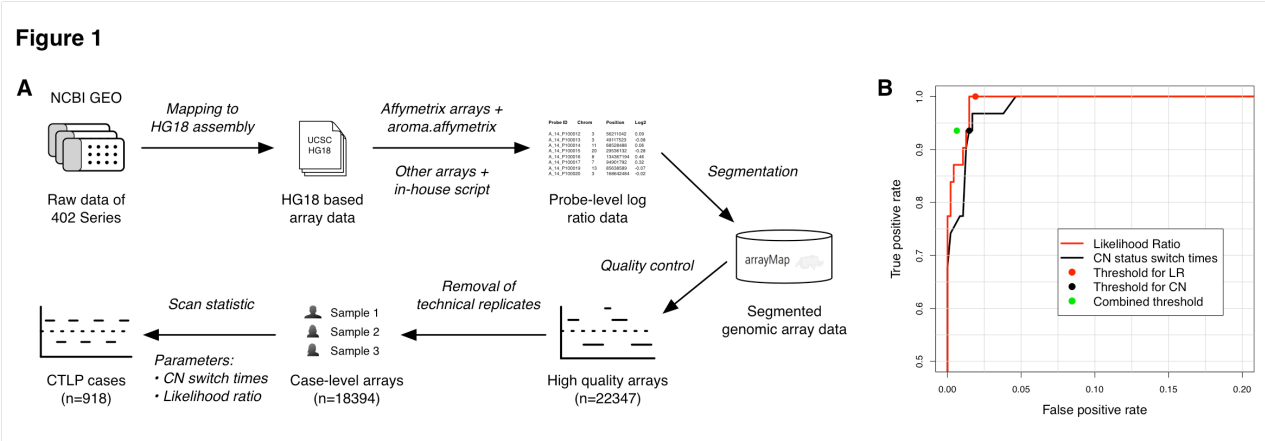


Figure 1 The ROC curve of the training set and selected thresholds. Two predictors were tested, copy number status change times and the likelihood ratio. Both predictors were integrated into the combined threshold.

Chromothripsis-like patterns across diverse tumor types

When evaluating the 1,269 CTLP events, we found a pronounced preference for some chromosomes; this preference showed only limited association with chromosome size (Fig. 2A). CTLP occurred more frequently in chromosome 17 than in any other chromosome. This observation is in accordance with data reporting an association between chromothripsis and *TP53* mutation in Sonic-Hedgehog medulloblastoma and acute myeloid leukemia (Rausch et al., 2012). *TP53* is located in the p arm of chromosome 17, and is involved in cell cycle control, genome maintenance and apoptosis (Forbes et al., 2011; Vogelstein and Levine, 2000). Our dataset showed *TP53* losses in

438 out of 918 (~48%) CTLP cases, compared to 3,274 out of 17,476 (~19%) cases in the non-CTLP group ($P < 2.2\times10^{-16}$; two-tailed Fisher's exact test; Supplemental Table 2). 45 of the 438 *TP53* deletions were part of a CTLP. Apart from its frequent involvement in different cancer types, *TP53* mutation seems to be a recurring and possibly associated event in chromothripsis formation. Other chromosomes with relatively high incidences of CTLP included chromosome 8, 11 and 12.

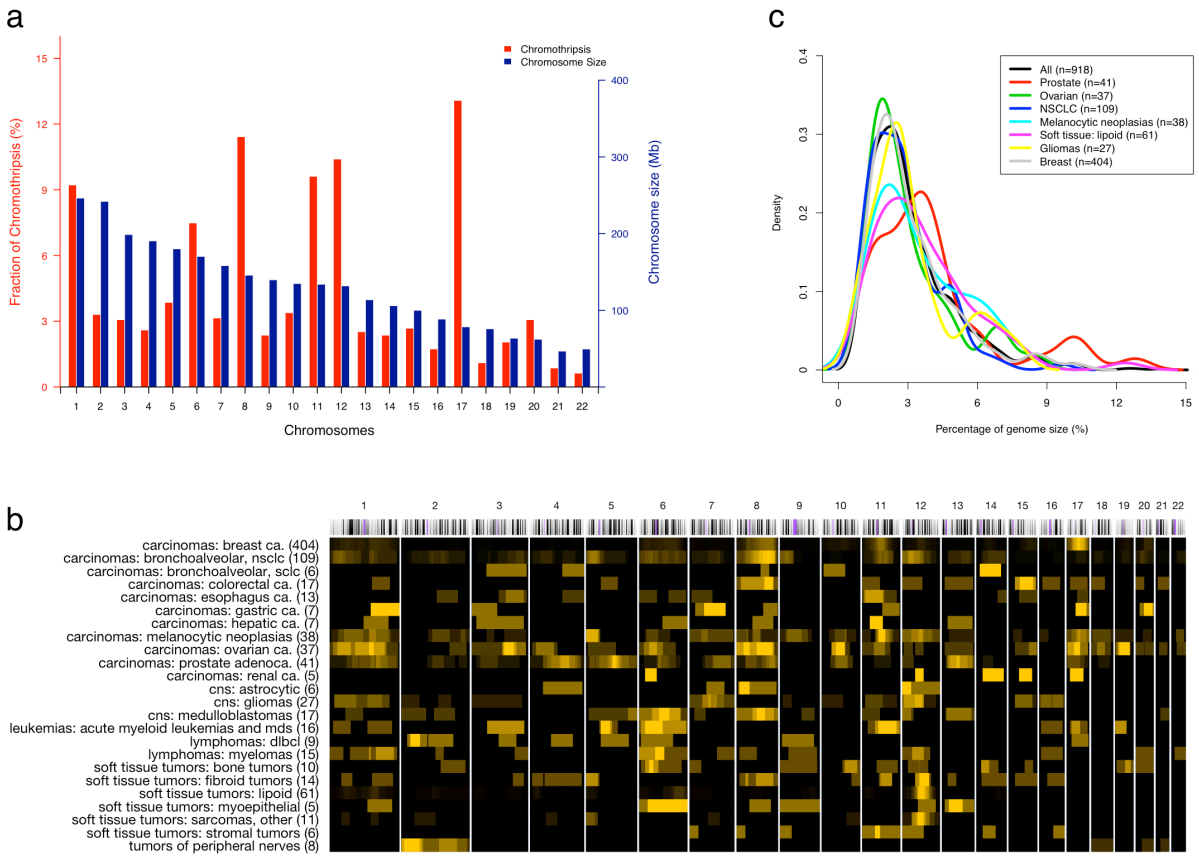


Figure 2 Frequency and coverage length distribution of CTLP in the genome. (a) Red and blue bars indicate CTLP frequency and chromosome size, respectively. (b) Distribution of CTLP regions among diagnostic groups. Each row represents a cancer type and each column represents a chromosome. We use a black-to-yellow gradient for representing CTLP frequencies ranging from lowest to highest, respectively, normalized for each row. The numbers in brackets indicate the number of cases. Groups with at least 5 CTLP cases are shown. (c) Sample derived fraction of genome involved in CTLP events. Density plots for common cancer types are shown.

In our study, genomic projection of regional CTLP frequencies revealed their heterogeneous distribution in different cancer types (Fig. 2B). The total length of fragmented genomic regions (CNA level and interspersed normal segments) accounted for 1%-14% of the corresponding genomes (Fig. 2C). The extent of our input data set with resulting high number of CTLP cases permitted an investigation of the frequency and genomic distribution of these patterns in different cancer types. Our input samples

represented 65 ”diagnostic groups”, as defined by combining ICD-O morphology and topography codes. The majority of samples (18,238) came from 50 diagnostic groups, each represented by more than 25 arrays. We observed in total of 918 CTLP events across all 18,394 cases, representing an overall ~5% prevalence. The 17 diagnostic groups represented by at least 45 cases, and having frequencies higher than 4% (CTLP high) are listed in Table 2 (full list in Supplemental Table 5).

Table 2 Frequency of chromothripsis-like patterns among cancer types

Cancer type	Chromothripsis-like cases				Input cases	Frequency (95% confidence interval)
	Oligo > 200K	Oligo ≤ 200K	BAC or cDNA	Total		
Soft tissue tumors: lipoid	49	12	0	61	114	53.5% (44%-62.8%)
Soft tissue tumors: fibroid tumors	14	0	0	14	59	23.7% (14%-36.9%)
Soft tissue tumors: sarcomas, other	9	0	2	11	48	22.9% (12.5%-37.7%)
Carcinomas: breast ca.	247	99	58	404	3652	11.1% (10.1%-12.1%)
Carcinomas: esophagus ca.	13	0	0	13	135	9.6% (5.4%-16.2%)
Carcinomas: bronchoalveolar, NSCLC	78	29	2	109	1164	9.4% (7.8%-11.2%)
Soft tissue tumors: bone tumors	7	3	0	10	123	8.1% (4.2%-14.8%)
Carcinomas: bronchoalveolar, SCLC	3	3	0	6	90	6.7% (2.7%-14.5%)
Carcinomas: prostate adenoca.	1	40	0	41	653	6.3% (4.6%-8.5%)
CNS: CNS PNET	4	0	0	4	65	6.2% (2%-15.8%)
Carcinomas: melanocytic neoplasias	31	1	6	38	621	6.1% (4.4%-8.4%)
Soft tissue tumors: myoepithelial	3	0	2	5	85	5.9% (2.2%-13.8%)
Carcinomas: ovarian ca.	31	5	1	37	801	4.6% (3.3%-6.4%)
Carcinomas: gastric ca.	1	5	1	7	160	4.4% (1.9%-9.2%)
CNS: gliomas	14	13	0	27	669	4% (2.7%-5.9%)
Soft tissue tumors: stromal tumors	5	1	0	6	151	4% (1.6%-8.8%)
CNS: medulloblastomas	13	4	0	17	430	4% (2.4%-6.4%)

Only cancer types with input cases ≥ 45 and frequency ≥ 4% are shown

The original study hypothesized that chromothripsis has a high incidence in bone tumors (Stephens et al., 2011). Notably, several soft tissue tumor types appeared in our ”CTLP high” frequency set (6 out of 17), including the 3 types with the highest scores. Moreover, the high prevalence of CTLP in soft tissue tumors was reflected in the ICD-O specific frequencies (Supplemental Table 6). The genesis and/or effect of multiple localized chromosomal breakage-fusion events may be related to specific molecular mechanisms in those tumor types. Notably, gene fusions are well-documented recurring events in sarcomas (Mitelman et al., 2007), in contrast to most other solid tumors. So far, more than 40 fusion genes have been recognized in sarcomas and treated as potential diagnostic and prognostic markers (Mitelman et al., 2007). Possibly, the double-strand breaks and

random fragment stitching events in chromothripsis are liable to generate oncogenic fusion genes (Stephens et al., 2011). Further sequencing-based efforts will be needed to identify the true extent of fusion gene generation and to elucidate their functional impact in chromothripsis cases.

Genomic context of chromothripsis-like events

It has been hypothesized that chromothripsis is a one-off cellular crisis generating a malignant clone in a very short time (Maher and Wilson, 2012; Stephens et al., 2011). In many of the CTLP samples in our study, highly fragmented chromosomal regions were embedded in larger CNA regions showing variations in patterns and overall extent (Fig. 3A). To test whether chromothripsis events are associated with overall genomic instability, we examined the extent of all copy number imbalances detected in our dataset. Comparing the 918 CTLP positive arrays with the remainder of 17,476 CTLP negative arrays, we found that CTLP samples tended to have higher proportions of CNA extent in their genomes ($P < 2.2\times10^{-16}$; Kolmogorov-Smirnov test) (Fig. 3B,C). This indicated that chromothripsis events frequently co-occur with other types of copy number aberrations. Plausible and non-exclusive explanations could be that chromothripsis might frequently arise due to previously established errors in the maintenance of genomic stability, or that chromothriptic aberrations involving genomic maintenance genes may predispose to the acquisition of additional CNA. For those frequent cases exhibiting additional non-chromothripsis CNA events, their possible contribution to oncogenesis has to be considered when modeling the role of chromothripsis in cancer development.

Potential mechanisms for chromosome shattering

While the mechanism(s) responsible for the generation of chromothripsis remain elusive, a number of studies have proposed hypotheses including ionizing radiation (Stephens et al., 2011), DNA replication stress (Liu et al., 2011), breakage-fusion-bridge cycles (Rausch et al., 2012; Stephens et al., 2011), premature chromosome compaction (Meyerson and Pellman, 2011), failed apoptosis (Fullwood et al., 2011; Tubio, 2011) and micronuclei formation (Crasta et al., 2012). In our dataset, although most (76%) CTLP cases presented single chromosome CTLP events, in approximately 24% CTLP cases affected

at least 2 chromosomes (Fig. 4A). For certain candidate mechanisms, e.g. micro-nucleus formation due to mitotic delay, this observation would imply more than one event whereas the observation is appears compatible with an aborted apoptosis process.

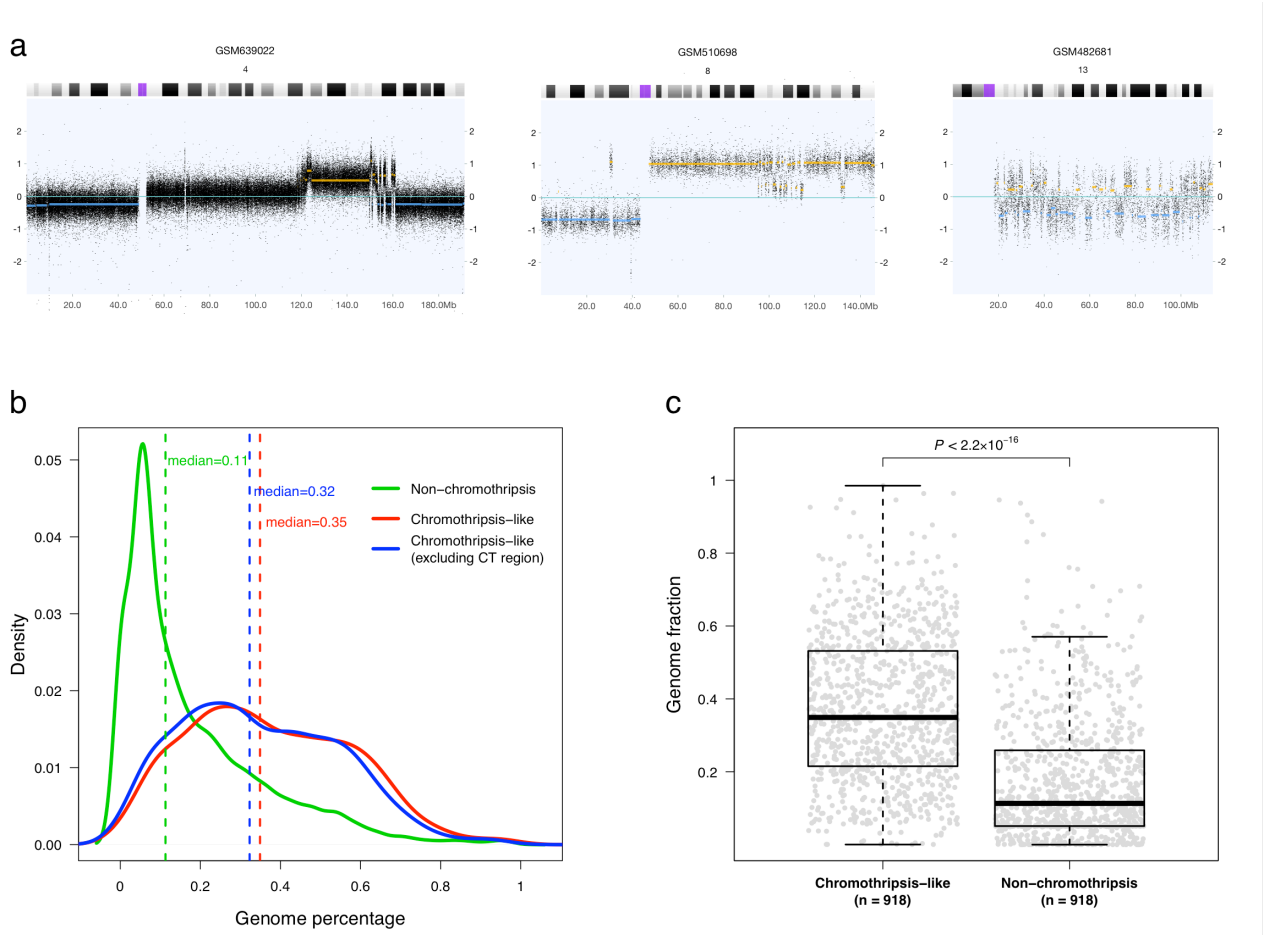


Figure 3 Examples for the genomic context of CTLP events. (a) Copy number profiles of the relevant chromosomes identified to have changes suggestive of chromothripsis. Chromosomal fragmentation events are embedded in other types of copy number aberrations, and exhibit different combination patterns. The x-axis indicates genomic locations in Mb, and the y-axis is the log2 value of probe signal intensity. Yellow and blue lines represent genomic gains and losses respectively. (b) Distribution of CNAs account for the percentage of the genome between CTLP and non-CTLP cases. CT, chromothripsis. (c) Total length of CNAs per sample. Each point represents one sample. For the non-CTLP group, 918 samples were randomly chosen from the total sample of 17,476 cases to generate an equally sized comparison. *P*-value, indicating significant difference between the genome fraction distributions of two groups, is based on Kolmogorov-Smirnov test.

In order to gain insight into the potential mechanisms, we subsequently normalized the size of the detected CTLP regions relative to their respective chromosomes. The affected regions were classified into the categories ”arm-level” (≥ 90% arm length), ”chromosome-level” (≥ 80% chromosome length) or ”localized” (Fig. 4B). Arm-level CTLP events were observed with a relatively high frequency (~19%). In the arm-level patterns, the CTLP rearrangements were concentrated in one chromosome arm with the other arm of the same chromosome remaining normal or showing isolated CNA. Notably, one model that

closely conforms to this pattern involves breakage-fusion-bridge cycles (Artandi and Depinho, 2010; Holland and Cleveland, 2012; McClintock, 1938, 1941; Meyerson and Pellman, 2011; Rausch et al., 2012; Stephens et al., 2011). In general, such cycles start with telomere loss and end-to-end chromosome fusions. When the dicentric chromosomes are formed and pulled to opposite poles during anaphase, a double-strand DNA breakage occurs and acts as a starting point for the next cycle. Chromosomal rearrangements would gradually accumulate during the additional cycles, and should be concentrated in one chromosome arm. To explore this hypothesis, we investigated the proportion of CTLP regions extending to telomeres. Here, up to 44% of all CTLP chromosomes involved telomere regions. We performed simulations to explore whether this telomere region enrichment could be explained by chance. In brief, for each sample, we retained the location of CTLP region in the genome and shuffled the telomere position of each chromosome while keeping the length of each chromosome constant. In contrast to the actual observations, our simulation did not result in telomeric CTLP enrichment (P < 0.0001; 10,000 simulations; see Methods). Hence, we require an explanation for the high proportion of arm-level pulverization and statistically significant telomere enrichment. CTLP generation through breakage-fusion-bridge cycles would be a viable candidate hypothesis.

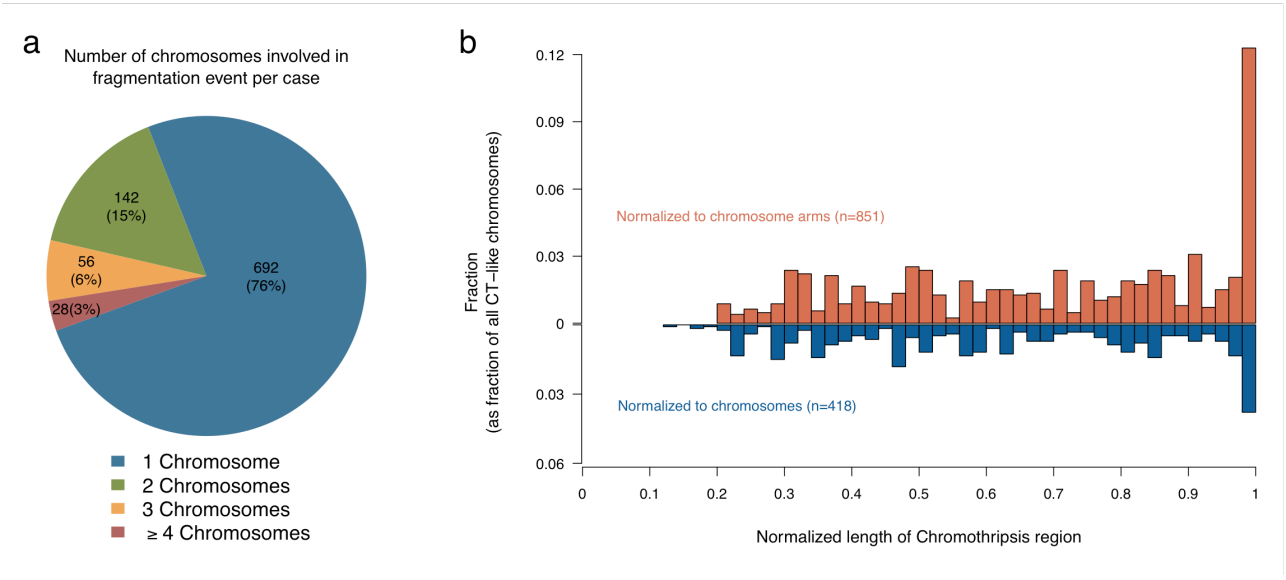


Figure 4 The distribution of CTLP regions in terms of chromosome number and length. (a) The number of chromosomes affected by CTLP per sample. The numbers outside and inside the brackets are number and percentage of CTLP samples respectively. (b) Length distribution of CTLP regions normalized to chromosome or chromosome arm lengths. For each chromosome, regions restricted on a single arm were normalized to arm lengths (red bars), otherwise were normalized to chromosome lengths (blue bars).

Clinical implications

Based on analyses of the clinical impact of chromothripsis patterns (Forment et al., 2012; Magrangeas et al., 2011; Molenaar et al., 2012), it has been claimed that these events may correlate with a poor outcome in the context of the respective tumor type. In our meta-analysis, we explored a general association of CTLP with clinical parameters, across the wide range of cancer entities reflected in our input data set. Clinical data was collected from GEO and from the publications of the respective series (Supplemental Tables 2, 7) and only parameters available for at least 1,000 cases were considered. From our dataset, CTLP seemed to occur at a more advanced patient age as compared to non-CTLP samples (Fig. 5A). CTLP mainly occurred at stage II and III (70%), which was significantly different from the stage distribution of total samples (55.2%) (P = 0.0149; Chi-square test) (Fig. 5B). No difference of grade distribution was observed in our dataset (P = 0.425; Chi-square test) where CTLP samples showed a predominance for grades 2 and 3, similar to the bulk of all samples (~80%). We also found that CTLP was overrepresented in cell lines compared to primary tumors ($P < 2.2\times10^{-16}$; two-tailed Fisher's exact test). For a subset of 1,203 patients, we were able to determine basic follow up parameters (time and death/ survival). 72 of these individuals showed CTLP in their tumor genomes. Notably, patients with CTLP survived a significantly shorter time than those without this phenomenon (P = 0.0039; log-rank test; Fig. 5C). Note that this analysis was based on a sample of convenience averaged over cancer and stage, etc. If we break down this dataset by cancer type, the numbers are not large enough to provide statistical confidence (Supplemental Fig. 4). While the cancer type independent association of CTLP patterns and poor outcome is intriguing, potential clinical effects of chromothripsis induced genome disruption should be evaluated in large and clinically more homogeneous data sets.

Sensitivity of array platforms for chromothripsis detection

Chromothripsis patterns have been reported from genomic datasets generated through different array and sequencing based techniques. We performed an analysis of the platform distribution of our CTLP samples, to estimate the possibility of platform associated detection bias. As the resolution of a platform depends both on type and density of the probes on an array, we divided the platforms into 4 groups according to their probe numbers and techniques (BAC/P1, DNA/cDNA, oligonucleotide ≤ 200K and

oligonucleotide > 200K). Although CTLP patterns were detected by all types of genomic arrays, a higher fraction of CTLP samples was found using data from high resolution oligonucleotide arrays (Fig. 6), possibly due to increased sensitivity related to higher probe density. Indeed, when performing platform simulations, the sensitivity of CTLP detection improved with increasing probe numbers (Supplemental Figs. 5, 6; see Methods). According to these simulations, array platforms consisting of more than 250k probes should be preferred when screening for chromothripsis-like events. Since our analysis relied on a variety of array platforms, we assume that the overall prevalence of CTLP patterns in cancer is higher than our reported 5% of samples.

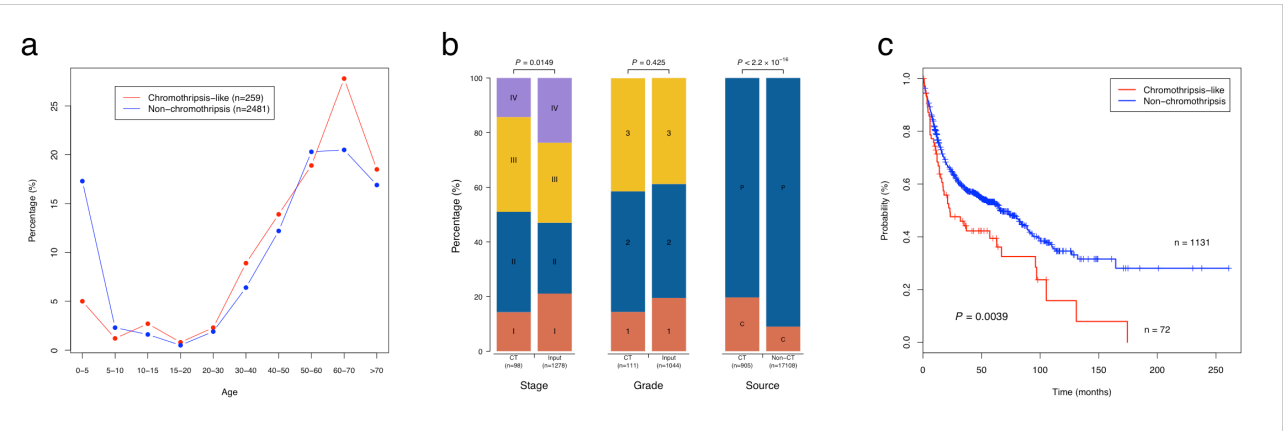


Figure 5 Clinical perspective on chromothripsis events. (a) Distribution of CTLP percentage versus patient age. (b) Distribution of sample stage, grade and source between CTLP and input dataset or non-CTLP cases. *P*-values are derived from Chi-square test (stage and grade) or two tailed Fisher's exact test (source). P, primary tumor; C, cell line. (c) Kaplan-Meier survival curves for CTLP versus non-CTLP cases. The *P*-value is based on log-rank test.

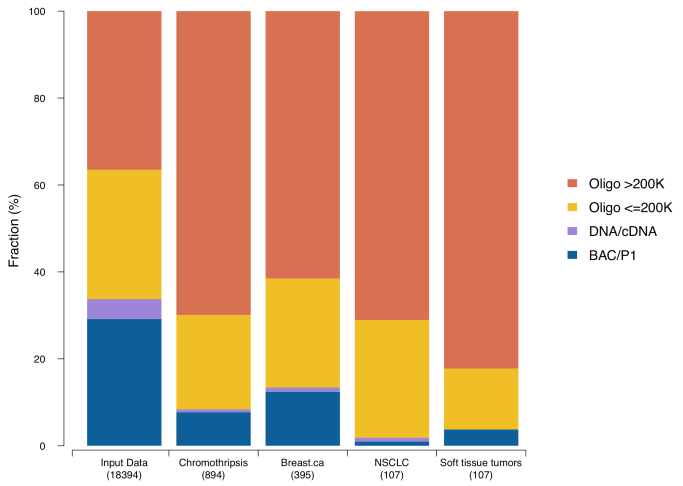


Figure 6 Platform distribution based on different resolutions and technology types. Different analysis groups are shown, including the whole input dataset, inferred CTLP cases and three cancer types. Oligo, oligonucleotide; NSCLC, Non-small cell lung cancer.

DISCUSSION

Chromothripsis has been characterized as a type of focally clustered genomic aberration events, occurring in a small subset of cancer genomes. In this study, we have identified 918 chromothripsis-like genome profiles, based on an analysis of copy number aberration patterns from 22,347 oncogenomic arrays, representing 132 cancer types. Despite the inherent limitations of such a meta-analysis approach, we were able to provide several new insights regarding the distribution of chromothripsis-like patterns and to produce a comprehensive estimate of CTLP incidence in a large range of cancer entities.

In our analysis, CTLP exhibited an uneven distribution along the tumor genomes, with disease related local enrichment. These ”CTLP dense” chromosomal regions may reveal associations between tumor type related cancer associated genes and molecular mechanisms behind chromothripsis events. This potential correlation is exemplified by the prevalence of mutant *TP53* in chromothriptic Li-Fraumeni syndrome associated Sonic-Hedgehog medulloblastomas (Rausch et al., 2012). As the extent of CTLP related deletions of the *TP53* locus indicates, chromothripsis based gene dosage changes may predispose to double-hit effects on specific tumor suppressors. In contrast, we found regional enrichment for CTLP with pre-dominant copy number gains on chromosomes 8, 11 and 12. In the initial study, chromosome 8 shattering was found in a small cell lung cancer cell line (Stephens et al., 2011). This event contained the *MYC* oncogene, which had be shown to be amplified in 10-20% of small cell lung cancers (Md and Md, 2008). Moreover, strong overexpression of *MYC* involved in a chromothripsis region was also detected in a neuroblastoma sample (Molenaar et al., 2012). In a study of colorectal tumors, chromosomes 8 and 11 were involved in concurrent pulverization events with generation of fusion genes, involving e.g. *SAPS3* and *ZFP91* (Bass et al., 2011). In a study on hepatocellular carcinoma, *CCND1* amplification was embedded within a chromothriptic event on chromosome 11 (Jiang et al., 2012). Therefore, the overall uneven distribution of CTLP may point to specific driver mutations that contribute to these events or to a class of chromothripsis derived cancer promoting mutations.

Among cancer types, we observed a high CTLP prevalence in a limited set of entities, particularly in among soft tissue tumors. This finding supports and improves upon a previous prediction of particularly high chromothripsis rate in bone tumors (Stephens et al.,

2011). Also, the uneven distribution of CTLP is a strong indicator for a disease related selection of specific genomic aberrations, supporting their involvement in the oncogenetic process.

In the initial study, the authors stated that chromothripsis could be a one-off cataclysmic event that generates multiple concurrent mutations and rearrangements (Stephens et al., 2011). Strikingly and in possible contradiction to this proposed singular ”shortcut” to cancer genome generation, we observed in the majority of CTLP samples additional, complex non-CTLP genome re-arrangements. Further efforts are needed to investigate the temporal order of chromothripsis and non-chromothripsis events observed in genomically complex samples.

Although the precise mechanisms that cause chromothripsis are unclear, several hypotheses have been put forward which might explain the phenomenon. Based on our observations, the number and uneven distribution of affected chromosomes in CTLP suggests that more than one mechanisms may be responsible for this catastrophic event. Furthermore, the normalized spatial distribution of shattered chromosomal regions, as well as the observed significant overlap between telomere and pulverized regions could be supportive of breakage-fusion-bridge cycles as one of the mechanisms behind this process.

We found that CTLP were related to overall advanced tumor stages and overall worse prognosis compared to non-CTLP cases. One possible explanation is that the numerous concurrent genetic alterations induced by chromothripsis events disturb a large number of genes and contribute to aggressive tumor phenotypes. By themselves, these observations do not explain if chromothripsis is an early event promoting aggressive tumor behavior with fast growth rates and reduced response rates to therapeutic interventions; or if this observation relates to underlying primary mutations predisposing to genomic instability, aggressive clinical behavior and CTLP as an epiphenomenon. Interestingly, the high rate of *TP53* involvement by itself would support both possibilities for this gene, i.e. chromothripsis as result of *TP53* mutation as well as chromothriptic events with *TP53* locus involvement promoting an aggressive clinical behavior.

In conclusion, CTLP represent a striking feature occurring in a limited set of cancer genomes, and can reliably be detected using biostatistical methods. The observed patterns may reflect on heterogenous biological phenomena beyond a single class of ”chromothripsis” events, and probably vary in their specific impact on oncogenesis. Fragmentation hotspots derived from our large-scale data set may promote the detection of markers involved in chromothriptic rearrangements, or may be used for assigning disease related effects to a CTLP induced genomic events.

Methods

Genome-wide microarrays and data preparation

In this study, we screened 402 GEO series (Barrett et al., 2013), including 22,347 high quality genomic arrays (Supplemental Table 2). All selected arrays were human cancer samples hybridized onto genome wide array platforms. The normalized probe intensities, segmented data and quality information were obtained from the arrayMap database, which is a publicly available reference database for copy number profiling data (Cai et al., 2012). In brief, the annotated data was obtained by the following processing pipeline: for Affymetrix arrays, the aroma.affymetrix R package was employed to generate log2 scale probe level data (Bengtsson et al., 2008); for non-Affymetrix arrays, available probe intensity files were processed; CBS (Circular Binary Segmentation) algorithm (Olshen et al., 2004) was performed to obtain segmented copy number data. The probe locations were mapped on the human reference genome (UCSC build hg18). In the case of technical repeats (e.g. one sample was hybridized on multiple platforms), only one of the arrays was considered for analysis (preferably with the highest resolution and/or best overall quality). The array profiling can be visualized through the arrayMap website.

Scan-statistic based chromothripsis pattern detection algorithm

To detect chromothripsis-like cases, we formulated an algorithm identifying clustering of copy number status changes in the genome. Several parameters were considered to define the alteration of copy number status:

i) The thresholds of log2 ratio for calling genomic gains and losses. These values were array specific and stored in arrayMap database. For each array, the thresholds were obtained from related publications or empirically assigned based on the log2 ratio distribution.

ii) The intensity distance between adjacent segments. Due to local correlation effects between probes or the existence of background noise, the segmentation profiles occasionally exhibit subtle striation patterns. This pattern is constituted with a large number of small segments, which is unlikely to be a biological phenomenon. To reduce artificial copy number status change, the distance of signal intensity between adjacent segments was used as a threshold, and defined here as the sum of the absolute values to call gains and losses. If the distance of two adjacent segments differed by less than this threshold, the copy number status change was not considered.

iii) Segment size. The resolution of a platform depends on the density of probes on the array. One of the platforms with the highest density in our dataset is Affymetrix SNP6, which contains 1.8 million polymorphic and non-polymorphic markers with the mean inter-marker distance of 1.7 kb. It provides a practical resolution of 10 to 20 kb. Therefore, in this study, segments smaller than 10 kb were removed.

In order to identify clustering of copy number status changes, a scan-statistic likelihood ratio based on the Poisson model was employed (Kulldorff, 1997). In our implementation, a fixed-size window was moved along the genome and for each window the likelihood ratio was computed from observed and expected copy number status change times. Let G be the genome represented linearly, and W is a window with fixed size. As the window W moves over G, it defines a collection of zones Z, where $Z \subset G$. Let n_W denotes the observed copy number status change times in window W, and n_G the total number of observed status change in G. μ_W is the expected status change times in window W, and is calculated as $W/G \times n_G$. The likelihood function is expressed as

$$\lambda = \begin{cases} \left(\frac{n_W}{\mu_W}\right)^{n_W} \left(\frac{n_G - n_W}{n_G - \mu_W}\right)^{n_G - n_W} & if \quad \frac{n_W}{\mu_W} > \frac{n_G - n_W}{n_G - \mu_W} \\ 1 & otherwise \end{cases}$$

This function detects the zone that is most likely to be a cluster.

Due to lack of a prior knowledge about the size of *W*, we predefined a series of window sizes from 30 Mb to 247 Mb (Supplemental Table 8), which were based on chromosome sizes. The scanning process was repeated for the series of window sizes for each sample. When *W* moved over *G*, the step length was set to 5 Mb, and there was no overlap between different chromosomes in window *W*. In this way, for each genome, the collection of *Z* contained 4,414 windows in various sizes. The window that maximized the likelihood ratio defined the most probable CTLP region. Thus it can detect both the location and the size of the cluster. When analyzing the complete input dataset, the window with the highest likelihood ratio was selected as a candidate of chromothripsis for each chromosome of the 22,347 arrays. The R script for detecting CTLP cases can be provided upon request.

Analysis of fragment enrichment in telomere region

Telomere positions were simulated to test the DNA fragment enrichment. For each case, the CTLP region was kept at its location in the genome. Locations of chromosome terminals were randomly selected while the length of each chromosome was kept. A genomic interval of 5 Mb from the chromosome terminal was considered as the telomere region. The simulation was performed 10,000 times.

Simulation of platform resolution

The 15 Affymetrix SNP6 CTLP chromosomes in the training set were used for simulation. For each genome, a certain number of probes were randomly chosen from the original probe set. These probes generally represented the profile that the same sample was hybridized on a platform with corresponding resolution. Then the CTLP pattern detection

algorithm was applied on the simulated arrays, and the number of cases that passed the thresholds were recorded.

Statistical testing

The significance in the number of CTLP cases with *TP53* loss in comparison to those in non-CTLP cases was assessed using two-tailed Fisher's exact test. We performed a Kolmogorov-Smirnov test to compare the distributions of copy number aberration proportions in the genome between CTLP and the other cases. The Chi-square test was used to assess the significance in the distribution of both patient stage and grade in CTLP and the whole input dataset. The associations between the number of cell lines in CTLP and non-CTLP cases were tested by two-tailed Fisher's exact test. The difference in the survival curve between two subgroups was evaluated by log-rank test.

Acknowledgments

The authors would like to thank Henrik Bengtsson and Ni Ai for their help with the project. This work was supported by grants from University of Zurich, University Research Priority Program for Systems Biology/Functional Genomics. H.C is supported by China Scholarship Council.

Please address correspondence to:
Michael Baudis
Institute of Molecular Life Sciences, University of Zurich
michael.baudis@imls.uzh.ch
Mark D. Robinson
Institute of Molecular Life Sciences, University of Zurich
mark.robinson@imls.uzh.ch

References

Albertson, D. G., Collins, C., McCormick, F., and Gray, J. W., 2003. Chromosome aberrations in solid tumors. *Nat Genet*, **34**(4):369–76.

Artandi, S. E. and Depinho, R. A., 2010. Telomeres and telomerase in cancer. *Carcinogenesis*, **31**(1):9–18.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., *et al.*, 2013. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Research*, **41**(D1):D991–D995.

Bass, A. J., Lawrence, M. S., Brace, L. E., Ramos, A. H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A., *et al.*, 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent vti1a-tcf7l2 fusion. *Nat Genet*, **43**(10):964–968.

Bengtsson, H., Simpson, K., Bullard, J., and Hansen, K., 2008. aroma.affymetrix: A genetic framework in R for analyzing small to very large affymetrix data sets in bounded memory. *Tech Report #745 Department of Statistics, University of California, Berkeley*, .

Berger, M. F., Hodis, E., Heffernan, T. P., Deribe, Y. L., Lawrence, M. S., Protopopov, A., Ivanova, E., Watson, I. R., Nickerson, E., Ghosh, P., *et al.*, 2012. Melanoma genome sequencing reveals frequent prex2 mutations. *Nature*, **485**(7399):502–506.

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J. S., Dobson, J., Urashima, M., *et al.*, 2031. The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**(7283):899–905.

Cai, H., Kumar, N., and Baudis, M., 2012. arraymap: A reference resource for genomic copy number imbalances in human malignancies. *PLoS ONE*, **7**(5):e36944.

Chen, J.-M., Cooper, D. N.,Férec, C., Kehrer-Sawatzki, H., and Patrinos, G. P., 2010. Genomic rearrangements in inherited disease and cancer. *Seminars in Cancer Biology*, **20**(4):222–233.

Chiang, C., Jacobsen, J. C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R. E., Kirby, A., Lindgren, A. M., Rudiger, S. R., *et al.*, 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet*, **44**(4):390–397.

Chin, L. and Gray, J. W., 2008. Translating insights from the cancer genome into clinical practice. *Nature*, **452**(7187):553–563.

Crasta, K., Ganem, N. J., Dagher, R., Lantermann, A. B., Ivanova, E. V., Pan, Y., Nezi, L., Protopopov, A., Chowdhury, D., and Pellman, D., *et al.*, 2012. Dna breaks and chromosome pulverization from errors in mitosis. *Nature*, **482**(7383):53–58.

Deakin, J. E., Bender, H. S., Pearse, A.-M., Rens, W., O'brien, P. C. M., Ferguson-Smith, M. A., Cheng, Y., Morris, K., Taylor, R., Stuart, A., *et al.*, 2012. Genomic restructuring in the tasmanian devil facial tumour: Chromosome painting and gene mapping provide clues to evolution of a transmissible tumour. *PLoS Genet*, **8**(2):e1002483.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., *et al.*, 2011. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, **39**(Database):D945–D950.

Forment, J. V., Kaidi, A., and Jackson, S. P., 2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nature Reviews Cancer*, **12**(10):663–670.

Fullwood, M. J., Lee, J., Lin, L., Li, G., Huss, M., Ng, P., Sung, W.-K., and Shenolikar, S., 2011. Next-generation sequencing of apoptotic dna breakpoints reveals association with actively transcribed genes and gene translocations. *PLoS ONE*, **6**(11):e26054.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N. D., Kanchi, K. L., Maher, C. A., Fulton, R., Fulton, L., Wallis, J., *et al.*, 2012. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell*, **150**(6):1121–1134.

Hanahan, D. and Weinberg, R. A., 2011. Hallmarks of cancer: The next generation. *Cell*, **144**(5):646–674.

Holland, A. J. and Cleveland, D. W., 2012. Chromoanagenesis and cancer: mechanisms and con- sequences of localized, complex chromosomal rearrangements. *Nat Med*, **18**(11):1630–1638.

Jiang, Z., Jhunjhunwala, S., Liu, J., Haverty, P. M., Kennemer, M. I., Guan, Y., Lee, W., Carnevali, P., Stinson, J., Johnson, S., *et al.*, 2012. The effects of hepatitis b virus integration into the genomes of hepatocellular carcinoma patients. *Genome Research*, **22**(4):593–601.

Jones, D. T. W., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.-J., Pugh, T. J., Hovestadt, V., Stütz, A. M., *et al.*, 2012. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, **488**(7409):100–105.

Kim, T.-M., Xi, R., Luquette, L. J., Park, R. W., Johnson, M. D., and Park, P. J., 2012. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Research*, :1–40.

Kitada, K., Aida, S., and Aikawa, S., 2012. Coamplification of multiple regions of chromosome 2, including mycn, in a single patchwork amplicon in cancer cell lines. *Cytogenet Genome Res*, **136**(1):30–37.

Kitada, K., Taima, A., Ogasawara, K., Metsugi, S., and Aikawa, S., 2011. Chromosome-specific segmentation revealed by structural analysis of individually isolated chromosomes. *Genes Chromosom. Cancer*, **50**:217–27.

Kloosterman, W. P., Guryev, V., Roosmalen, M. V., Duran, K. J., Bruijn, E. D., Bakker, S. C. M., Letteboer, T., Nesselrooij, B. V., Hochstenbach, R., Poot, M., *et al.*, 2011a. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human Molecular Genetics*, **20**(10):1916–1924.

Kloosterman, W. P., Hoogstraat, M., Paling, O., Tavakoli-Yaraki, M., Renkens, I., Vermaat, J. S., van Roosmalen, M. J., van Lieshout, S., Nijman, I. J., Roessingh, W., *et al.*, 2011b. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome Biology*, **12**(10):R103.

Kloosterman, W. P., Tavakoli-Yaraki, M., Roosmalen, M. J. V., Binsbergen, E. V., Renkens, I., Duran, K., Ballarati, L., Vergult, S., Giardino, D., Hansson, K., *et al.*, 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded dna breaks and nonhomologous repair mechanisms. *Cell Reports*, **1**(6):648–655.

Kulldorff, M., 1997. A spatial scan statistic. *Commun statist*, **26**(6):1481–1496.

Lapuk, A. V., Wu, C., Wyatt, A. W., Mcpherson, A., Mcconeghy, B. J., Brahmbhatt, S., Mo, F., Zoubeidi, A., Anderson, S., Bell, R. H., *et al.*, 2012. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. *J. Pathol.*, **227**(3):286– 297.

Le, L. P., Nielsen, G. P., Rosenberg, A. E., Thomas, D., Batten, J. M., Deshpande, V., Schwab, J., Duan, Z., Xavier, R. J., Hornicek, F. J., *et al.*, 2011. Recurrent chromosomal copy number alterations in sporadic chordomas. *PLoS ONE*, **6**(5):e18846.

Liu, P., Erez, A., Nagamani, S. C. S., Dhar, S. U., Kołodziejska, K. E., Dharmadhikari, A. V., Cooper, M. L., Wiszniewska, J., Zhang, F., Withers, M. A., *et al.*, 2011. Chromosome

catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*, **146**(6):889–903.

Magrangeas, F., Avet-Loiseau, H., Munshi, N. C., and Minvielle, S., 2011. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. *Blood*, **118**(3):675–678.

Maher, C. A. and Wilson, R. K., 2012. Chromothripsis and human disease: Piecing together the shattering process. *Cell*, **148**(1-2):29–32.

McClintock, B., 1938. The production of homozygous deficient tissues with mutant characteristics by means of the aberrant mitotic behavior of ring-shaped chromosomes. *Genetics*, **23**:315–76.

McClintock, B., 1941. The stability of broken ends of chromosomes in zea mays. *Genetics*, **26**:234–82.

Md, T. S. and Md, G. K. D., 2008. Small cell lung cancer. *Mayo Clinic Proceedings*, **83**(3): 355– 367.

Meyerson, M. and Pellman, D., 2011. Cancer genomes evolve by pulverizing single chromosomes. *Cell*, **144**(1):9–10.

Mitelman, F., Johansson, B., and Mertens, F., 2007. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, **7**(4):233–245.

Molenaar, J. J., Koster, J., Zwijnenburg, D. A., van Sluis, P., Valentijn, L. J., van der Ploeg, I., Hamdi, M., van Nes, J., Westerman, B. A., van Arkel, J., *et al.*, 2012. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature*, **483**(7391):589–593.

Natrajan, R., Mackay, A., Lambros, M. B., Weigelt, B., Wilkerson, P. M., Manie, E., Grigoriadis, A., A'hern, R., Groep, P. V. D., Kozarewa, I., *et al.*, 2012. A whole-genome massively parallel sequencing analysis of brca1 mutant oestrogen receptor-negative and -positive breast cancers. *J. Pathol.*, **227**(1):29–41.

Naus, J. I., 1965. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, **60**:532–538.

Nik-Zainal, S., Loo, P. V., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., *et al.*, 2012. The life history of 21 breast cancers. *Cell*, **149**(5):994–1007.

Northcott, P. A., Shih, D. J. H., Peacock, J., Garzia, L., Morrissy, A. S., Zichner, T., Stuetz, A. M., Korshunov, A., Reimand, J., and Schumacher, S. E., *et al.*, 2012. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, **488**(7409):49–56.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M., 2004. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5**(4): 557–572.

Poaty, H., Coullin, P., Peko, J. F., Dessen, P., Diatta, A. L., Valent, A., Leguern, E., Prévot, S., Gombé-Mbalawa, C., Candelier, J.-J., *et al.*, 2012. Genome-wide high-resolution aCGH analysis of gestational choriocarcinomas. *PLoS ONE*, **7**(1):e29426.

Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P. A., *et al.*, 2012. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, **148**(1-2):59–71.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., *et al.*, 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**(1):27–40.

Stevens-Kroef, M., Weghuis, D. O., Croockewit, S., Derksen, L., Hooijer, J., Elidrissi-Zaynoun, N., Siepman, A., Simons, A., and Kessel, A. G. V., 2012. High detection rate of clinically relevant genomic abnormalities in plasma cells enriched from patients with multiple myeloma. *Genes Chromosom. Cancer*, **51**(11):997–1006.

Stratton, M. R., Campbell, P. J., and Futreal, P. A., 2009. The cancer genome. *Nature*, **458**(7239):719–724.

Tubio, X. E. J., 2011. When catastrophe strikes a cell. *Nature*, **24**:476–477.

Vogelstein, D. L. B. and Levine, A. J., 2000. Surfing the p53 network. *Nature*, **408**:307–10.

Wu, C., Wyatt, A. W., Mcpherson, A., Lin, D., Mcconeghy, B. J., Mo, F., Shukin, R., Lapuk, A. V., Jones, S. J. M., Zhao, Y., *et al.*, 2012. Poly-gene fusion transcripts and chromothripsis in prostate cancer. *Genes Chromosom. Cancer*, **51**(12):1144–1153.

Yates, L. R. and Campbell, P. J., 2012. Evolution of the cancer genome. *Nature Reviews Genetics*, **13**(11):795–806.

Zehentner, B. K., Hartmann, L., Johnson, K. R., Stephenson, C. F., Chapman, D. B., Baca, M. E. D., Wells, D. A., Loken, M. R., Tirtorahardjo, B., Gunn, S. R., *et al.*, 2012. Array-based karyotyping in plasma cell neoplasia after plasma cell enrichment increases detection of genomic aberrations. *American Journal of Clinical Pathology*, **138**(4):579–589.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S. L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., *et al.*, 2012. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, **481**(7380):157–163.

Supplementary Information

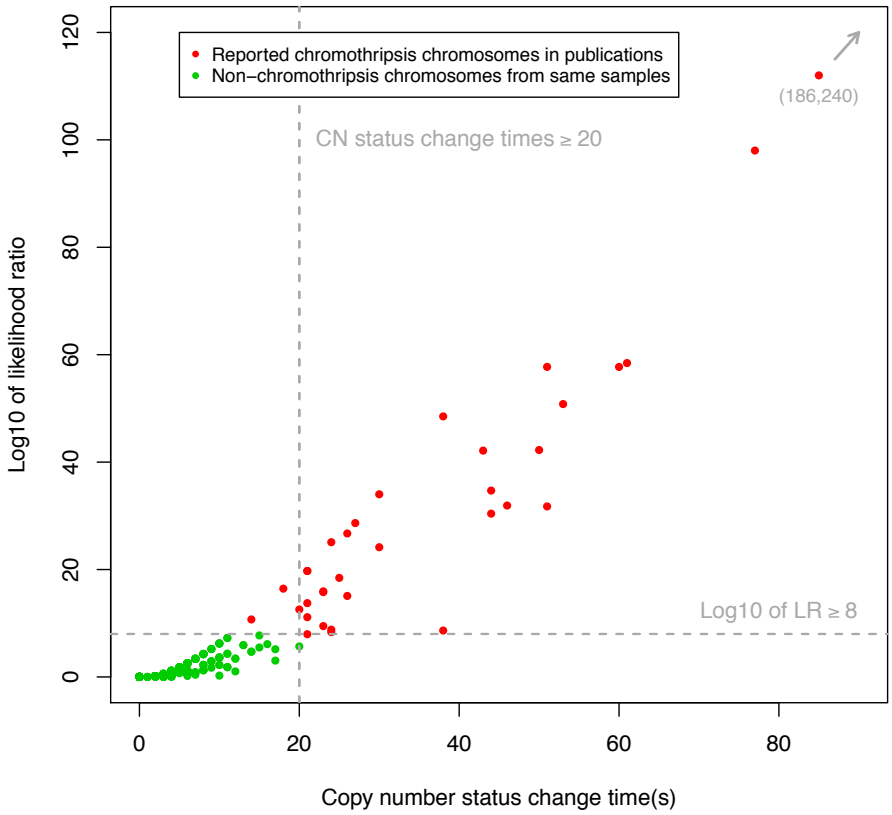
Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genomes

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C. Bagheri[3], Christian von Mering[1,2], Mark D. Robinson[1,2], and Michael Baudis[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2]Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

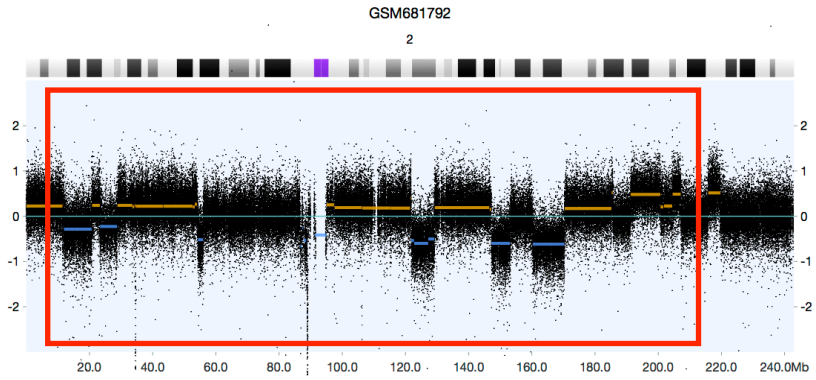
1. Supplementary Figures

2. Supplementary Tables

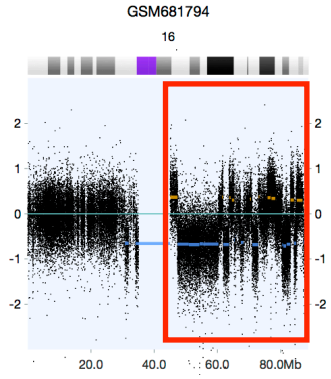


Supplementary Figure 1. Scatter plot of the training set. Copy number status change times compared to the likelihood ratio. Each point represents the window with the highest LR for each chromosome. The dashed lines indicate the selected thresholds. CN, copy number; LR, likelihood ratio.

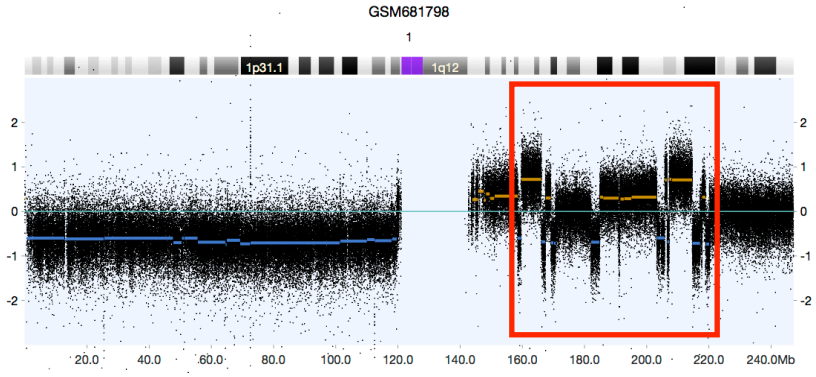
a



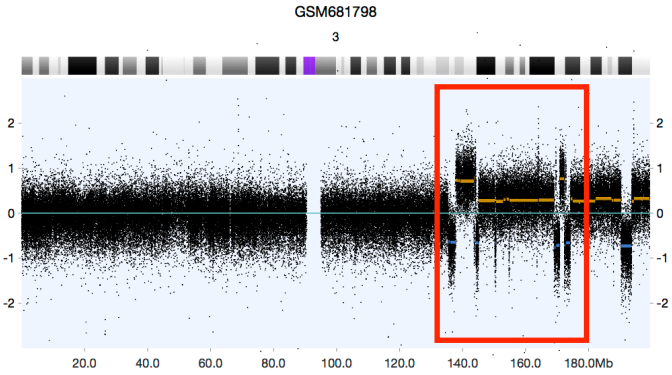
Log10(LR) = 8.6
Switch times = 38
Window size = 199.5 Mb



Log10(LR) = 31.9
Switch times = 46
Window size = 46.9 Mb

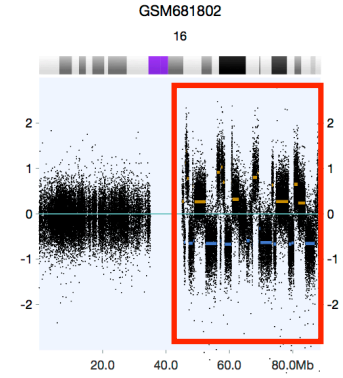


Log10(LR) = 15
Switch times = 26
Window size = 62.4 Mb

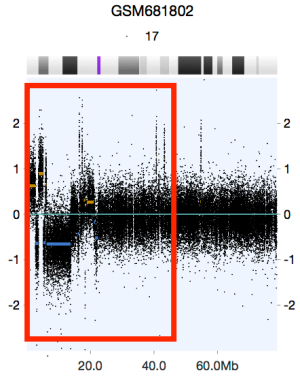


Log10(LR) = 24.2
Switch times = 30
Window size = 40 Mb

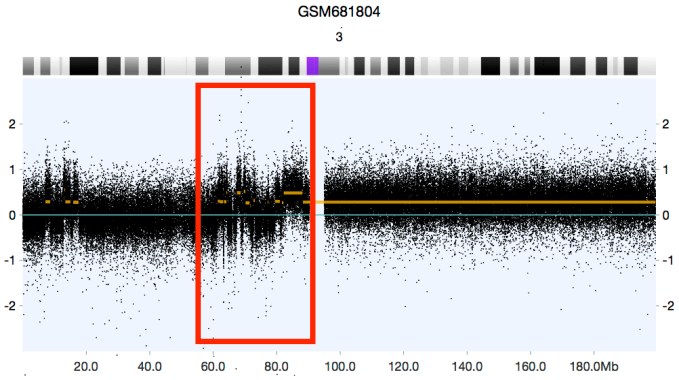
b



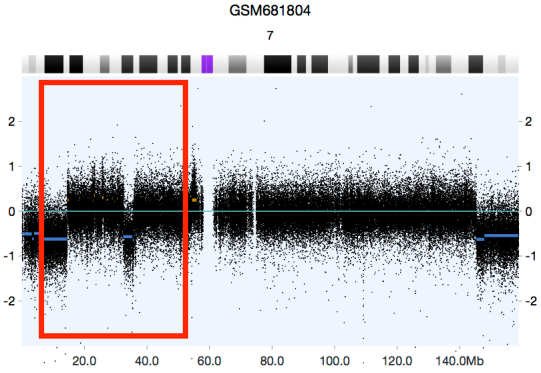
Log10(LR) = 42.1
Switch times = 43
Window size = 46.9 Mb



Log10(LR) = 15.9
Switch times = 23
Window size = 46.9 Mb

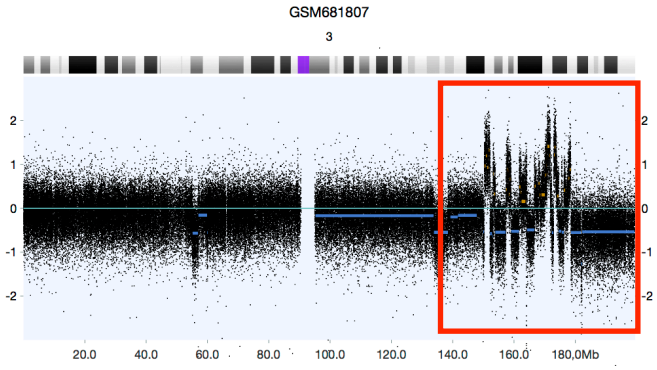


Log10(LR) = 19.7
Switch times = 21
Window size = 30 Mb

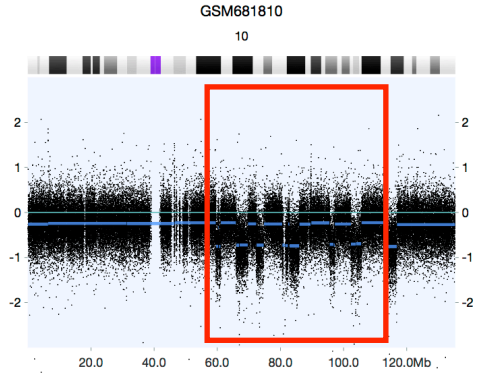


Log10(LR) = 19.7
Switch times = 21
Window size = 40 Mb

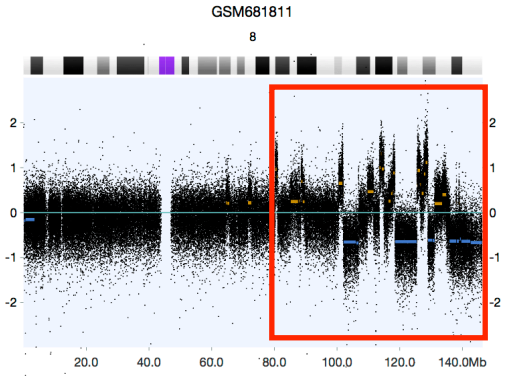
c



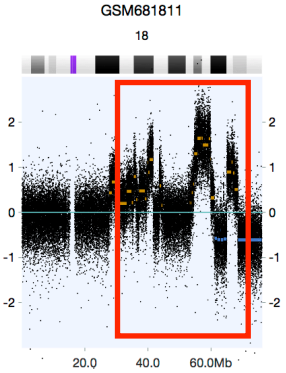
Log10(LR) = 58.4
Switch times = 61
Window size = 49.7 Mb



Log10(LR) = 13.7
Switch times = 21
Window size = 46.9 Mb

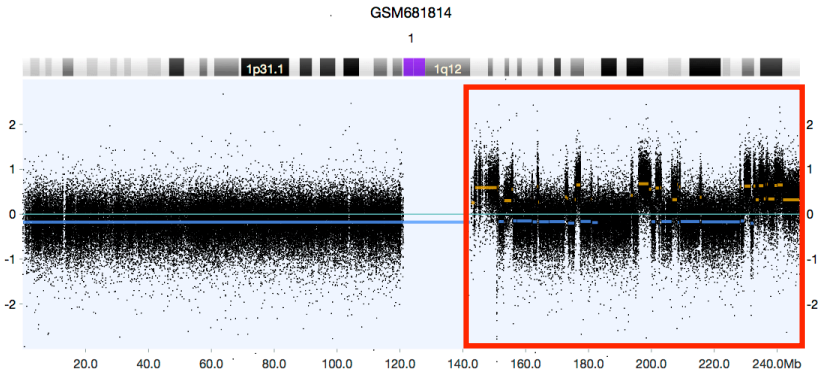


Log10(LR) = 57.7
Switch times = 60
Window size = 62.4 Mb

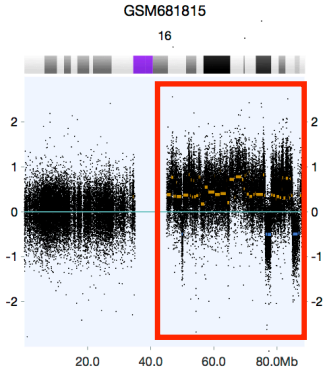


Log10(LR) = 26.7
Switch times = 26
Window size = 30 Mb

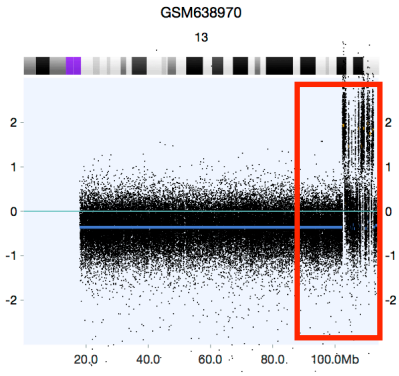
d



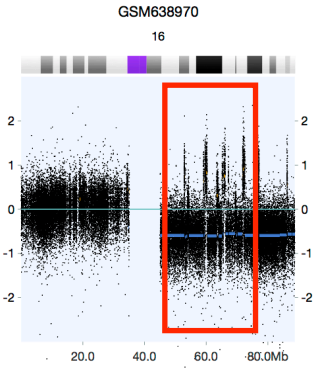
Log10(LR) = 50.8
Switch times = 53
Window size = 106.4 Mb



Log10(LR) = 98
Switch times = 77
Window size = 46.9 Mb

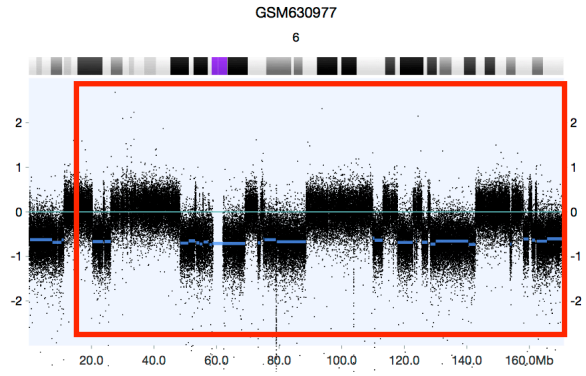


Log10(LR) = 42.3
Switch times = 50
Window size = 30 Mb

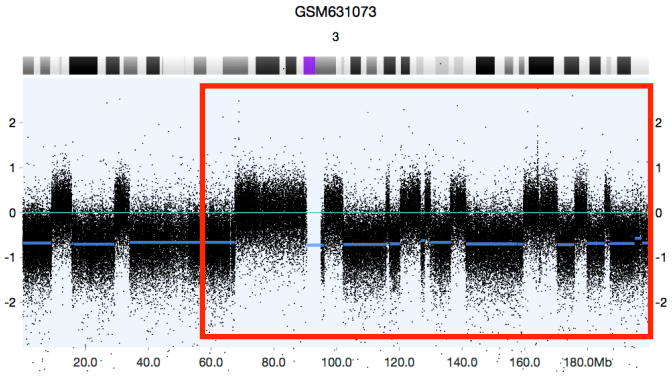


Log10(LR) = 34.7
Switch times = 44
Window size = 30 Mb

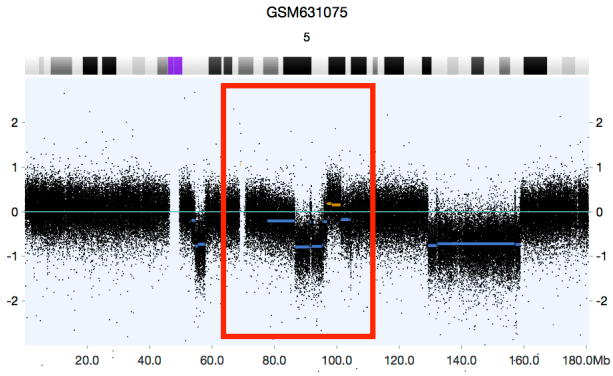
e



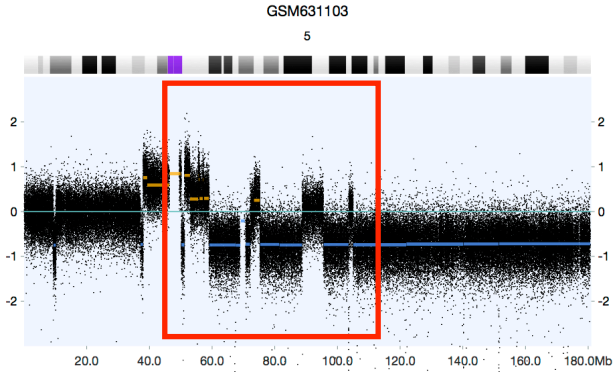
Log10(LR) = 31.8
Switch times = 51
Window size = 146.3 Mb



Log10(LR) = 8.8
Switch times = 24
Window size = 132.3 Mb

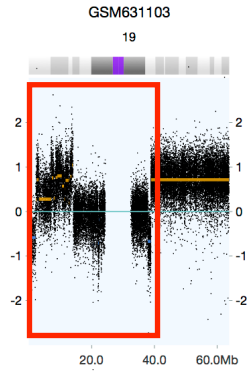


Log10(LR) = 10.7
Switch times = 14
Window size = 40 Mb

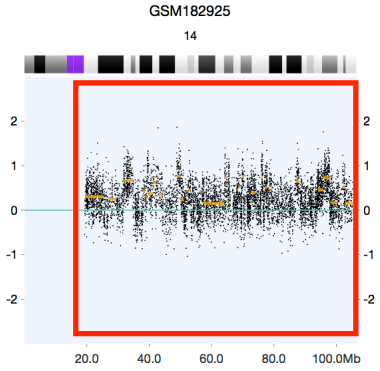


Log10(LR) = 18.4
Switch times = 25
Window size = 62.4 Mb

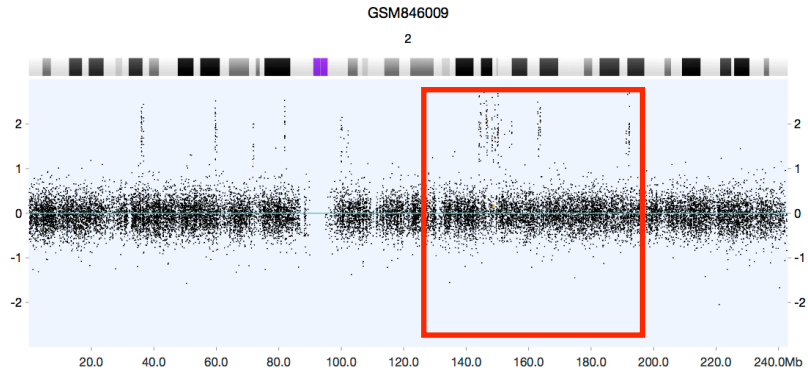
f



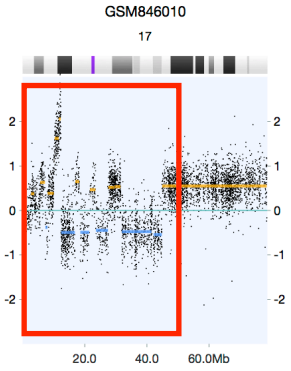
Log10(LR) = 28.7
Switch times = 27
Window size = 40 Mb



Log10(LR) = 57.7
Switch times = 51
Window size = 88.8 Mb

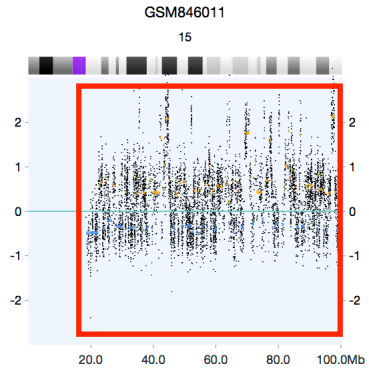


Log10(LR) = 34
Switch times = 30
Window size = 62.4 Mb

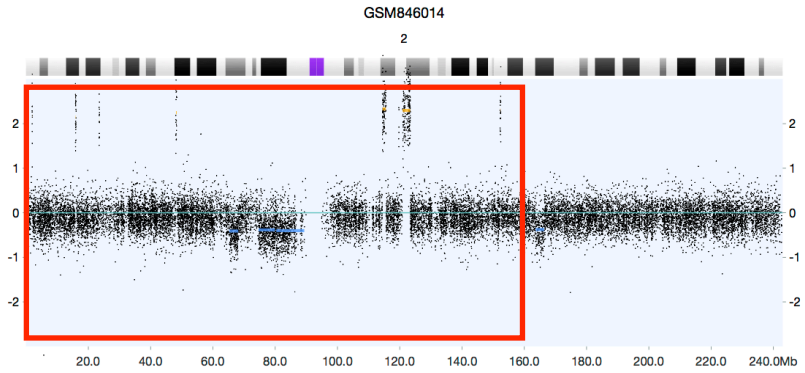


Log10(LR) = 16.4
Switch times = 18
Window size = 49.7 Mb

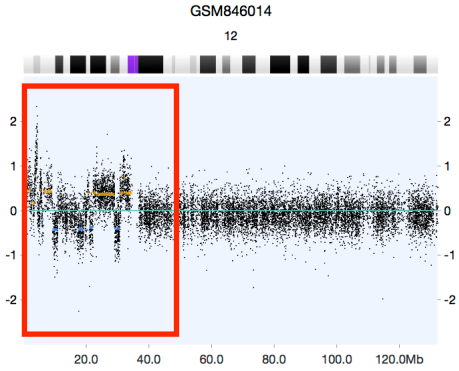
g



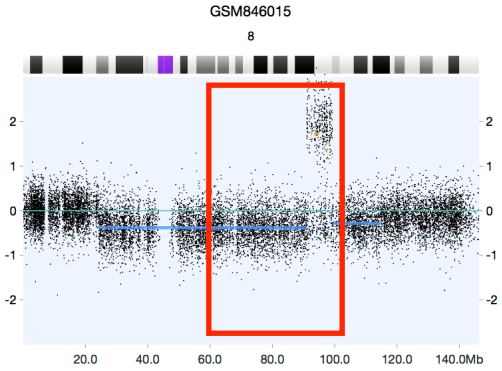
Log10(LR) = 240
Switch times = 186
Window size = 78.8 Mb



Log10(LR) = 11.1
Switch times = 21
Window size = 158.8 Mb

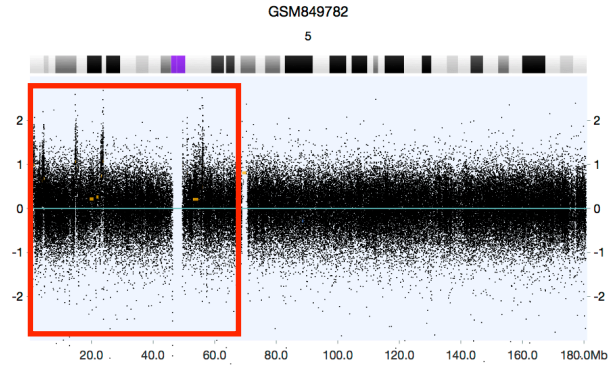


Log10(LR) = 25.1
Switch times = 24
Window size = 46.9 Mb

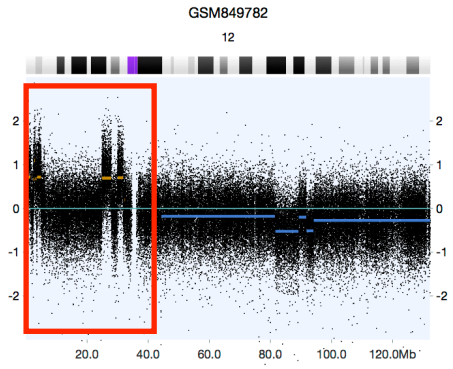


Log10(LR) = 48.5
Switch times = 38
Window size = 40 Mb

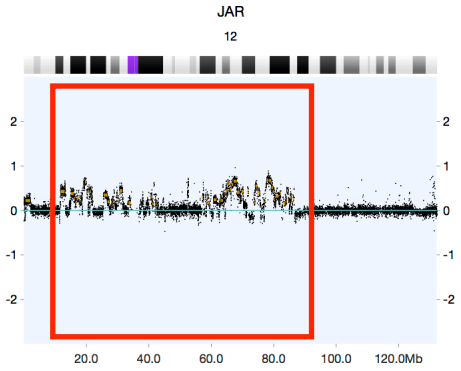
h



Log10(LR) = 30.4
Switch times = 44
Window size = 62.4 Mb

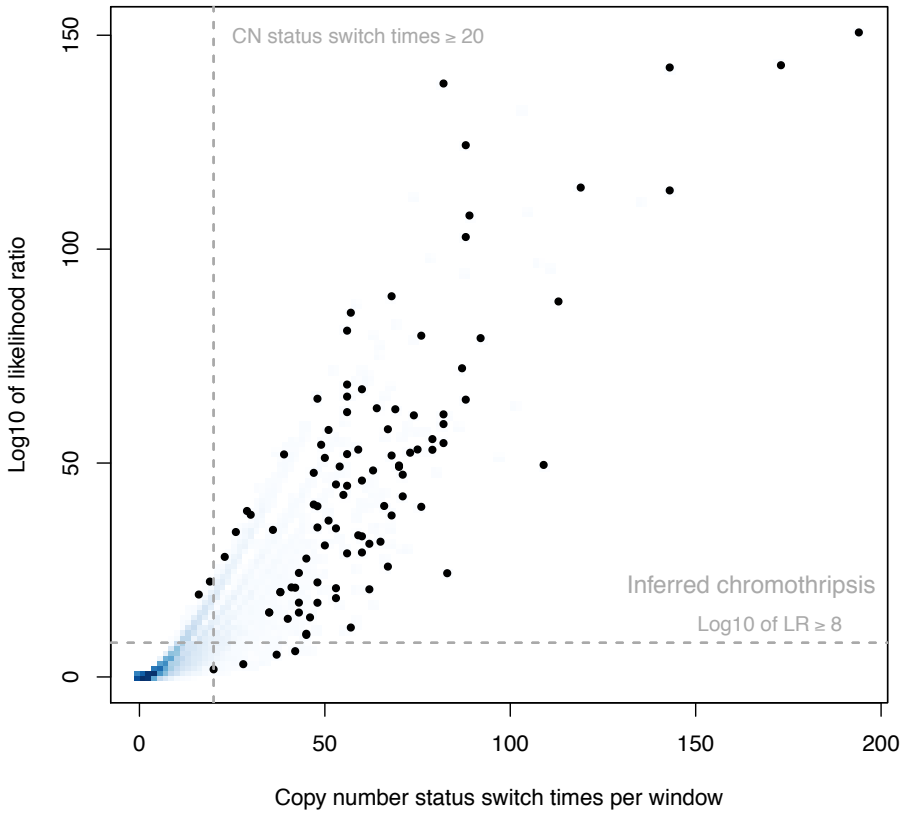


Log10(LR) = 12.6
Switch times = 20
Window size = 40 Mb

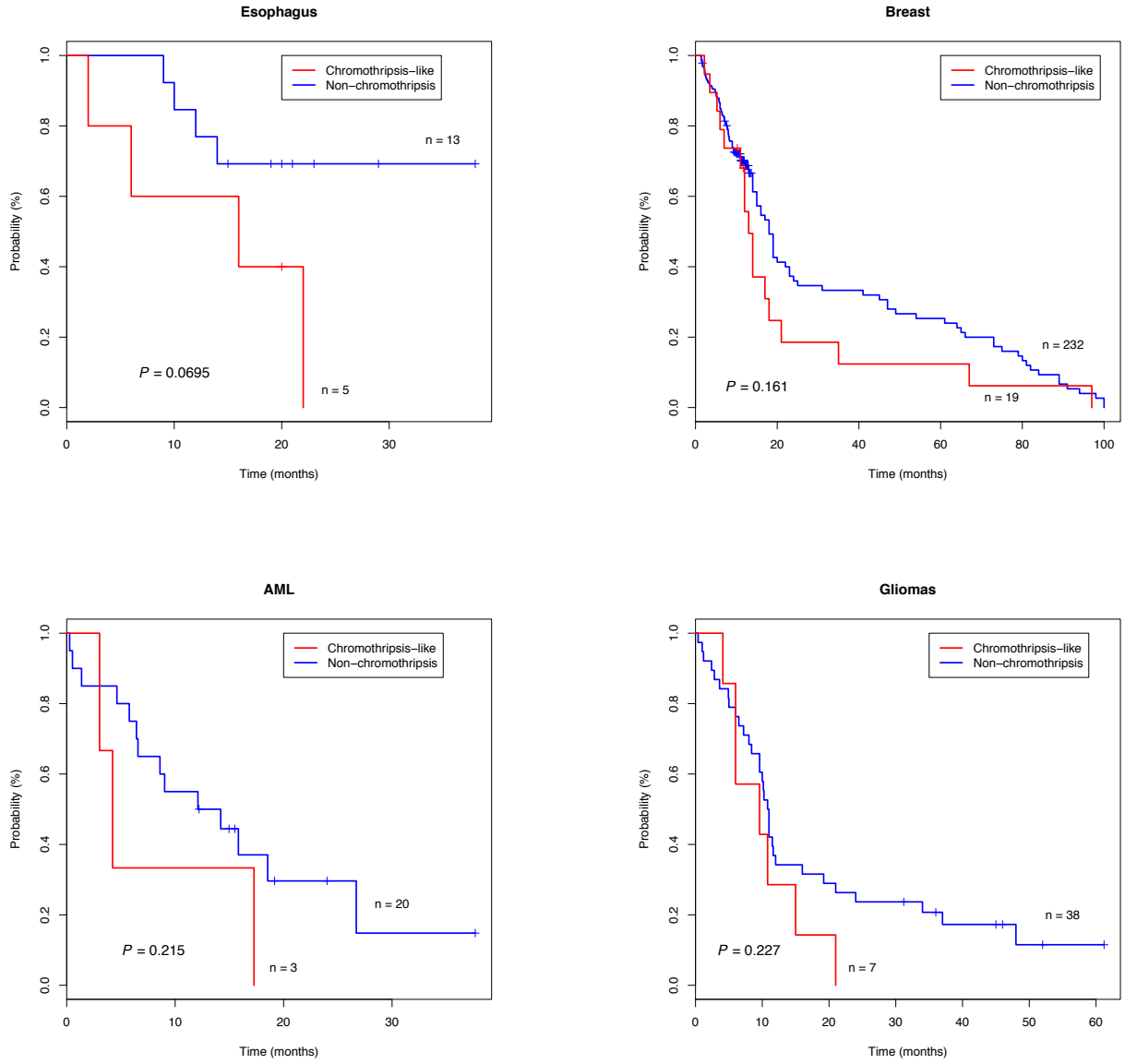


Log10(LR) = 107.8
Switch times = 149
Window size = 78.8 Mb

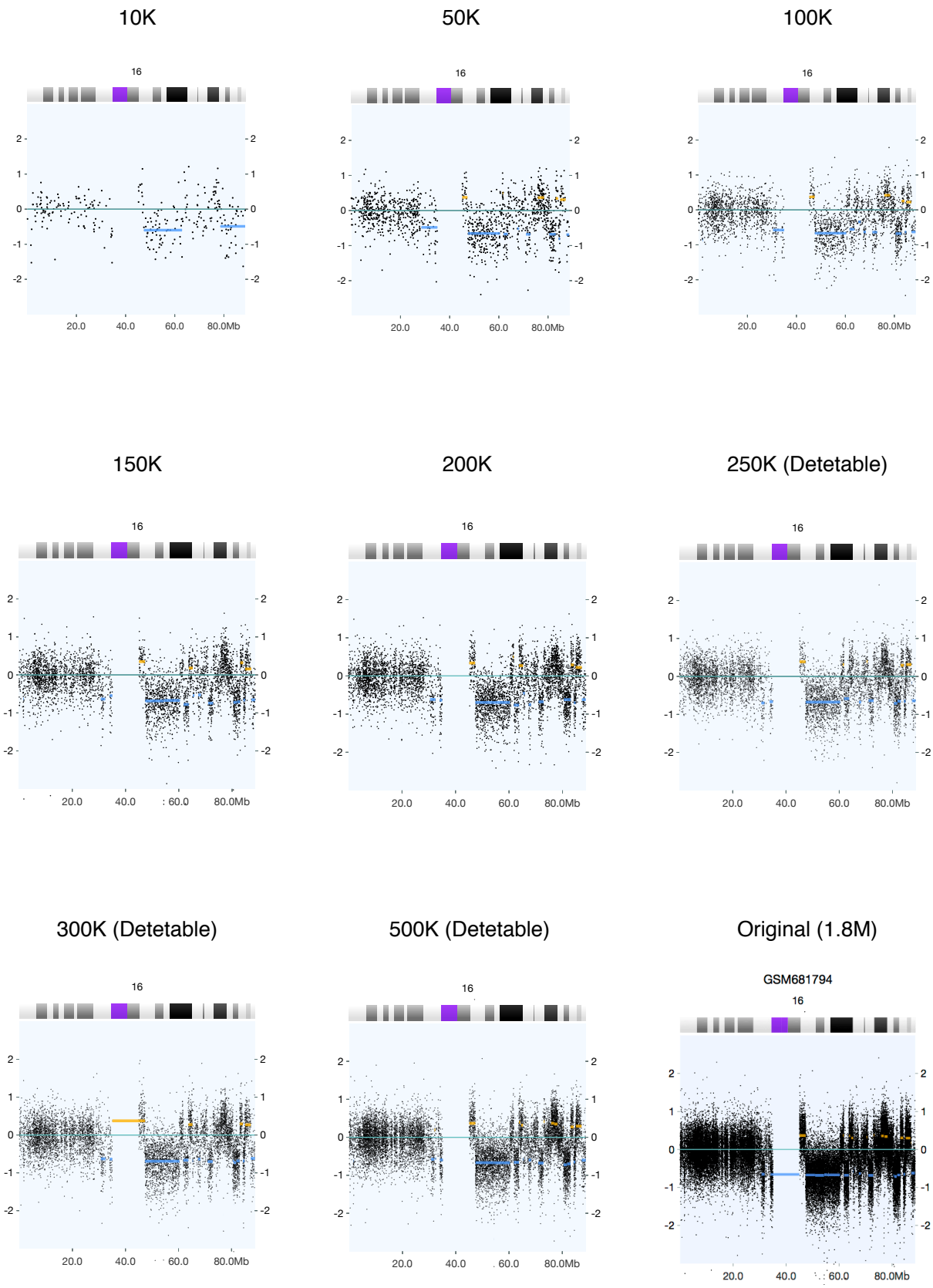
Supplementary Figure 2. The positive training set and CTLP detection algorithm performances. The red rectangles are chromothripsis-like regions identified by scan-statistic. For each plot, the parameters and corresponding values are shown in orange boxes. The schema of the chromosome is the same as in Figure 2.



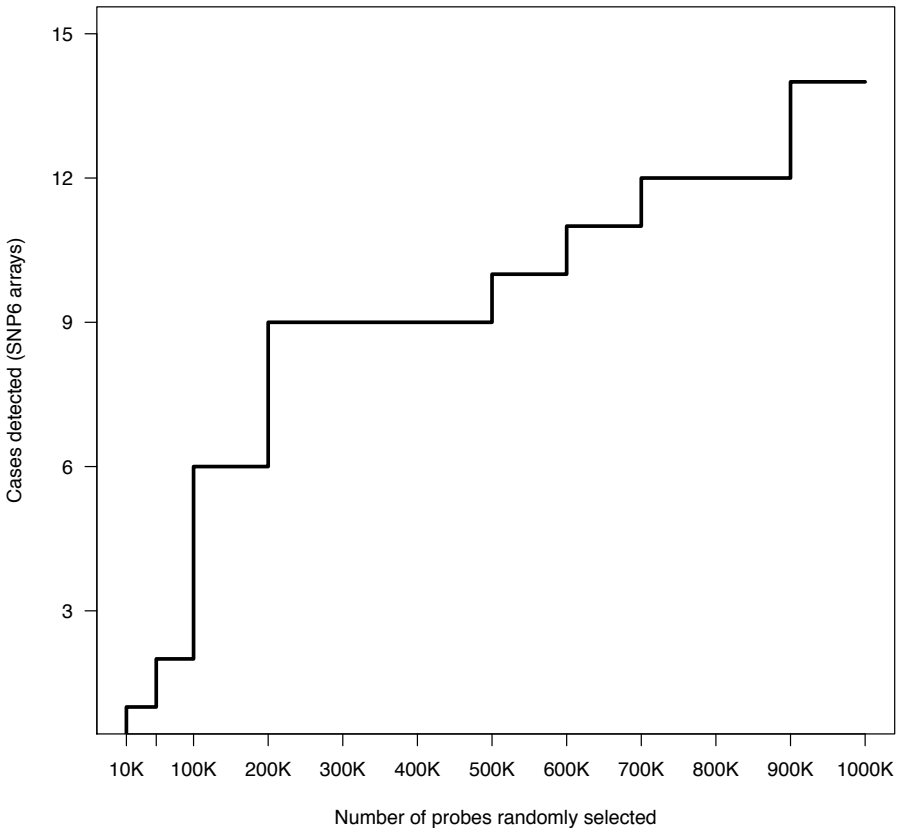
Supplementary Figure 3. Scatter plot of CTLP candidates. For each chromosome of the input dataset, the window with the highest likelihood ratio was considered as a CTLP candidate. The selected thresholds are indicated with dashed lines. The candidates falling in the upper right area are inferred CTLP cases.



Supplementary Figure 4. Kaplan-Meier survival curves for CTLP versus non-CTLP cases in specific cancer types.



Supplementary Figure 5. An example of the platform resolution based simulation from Affymetrix SNP6 array (1.8M). Chromosome 16 of GSM681794 is a reported chromothripsis event in multiple myeloma. The number above each plot represents the number of probes randomly selected from the original probe set. In this sample, the CTLP pattern is detectable starting at 250k probes.



Supplementary Figure 6. CTLP detection sensitivity of simulated platform resolutions. All the 15 chromothripsis chromosomes are from the positive training set analyzed using Affymetrix SNP6 platform.

Supplementary Table 1. Overview of input dataset

Category	Array-level	Case-level
Total number	22347	18394
Series	402	397
Platform	190	185
Cancer type (ICD-O)	132	132
Cancer type (diagnostic group)	65	65
Source (primary)	19623	16309
Source (cell line)	2309	1714
Source (relapse)	75	75
Source (metastasis)	340	296

Supplementary Table 7. Demographic and clinicopathologic characteristics of input and CTLP samples

Variable Name	Sample Number	Mean±SD/Median or %
Input set		
Male	1088	49.7%
Age(year)	2740	47.5 ± 25.7/55
AJCC Stage		
I	270	21.1%
II	331	25.9%
III	374	29.3%
IV	303	23.7%
Grade		
1	204	19.5%
2	435	41.7%
3	405	38.8%
Tumor recurrence	198	44.0%
Follow up (month)	1203	36.5 ± 34.6/26
Event		
Deceased	553	46.0%
Censored	650	54.0%
CTLP		
Male	46	32.6%
Age(year)	259	54.1 ± 20.7/59
AJCC Stage		
I	14	14.3%
II	36	36.7%
III	34	34.7%
IV	14	14.3%
Grade		
1	16	14.4%
2	49	44.1%
3	46	41.5%
Tumor recurrence	18	47.4%
Follow up (month)	72	32.9 ± 35.2/17.6
Event		
Deceased	46	63.9%
Censored	26	36.1%

† All information is based on the available clinical data

Supplementary Table 8. Sizes of sliding windows for the scan-statistic based algorithm

Size ID	Size (Mb)	Corresponding chromosome
1	247.249	Chromosome 1
2	242.951	Chromosome 2
3	199.501	Chromosome 3
4	191.273	Chromosome 4
5	180.857	Chromosome 5
6	170.899	Chromosome 6
7	158.821	Chromosome 7
8	146.274	Chromosome 8
9	140.273	Chromosome 9
10	135.374	Chromosome 10
11	134.452	Chromosome 11
12	132.349	Chromosome 12
13	114.142	Chromosome 13
14	106.368	Chromosome 14
15	100.338	Chromosome 15
16	88.827	Chromosome 16
17	78.774	Chromosome 17
18	76.117	Chromosome 18
19	63.811	Chromosome 19
20	62.435	Chromosome 20
21	46.944	Chromosome 21
22	49.691	Chromosome 22
23	40	NA
24	30	NA

Part III

DISCUSSION AND PERSPECTIVES

7 DISCUSSION

7.1 COPY NUMBER ABERRATION DATA IN HUMAN CANCER

Two databases for genomic copy number aberration in human cancer were launched and maintained by our group - arrayMap[23] and Progenetix[169]. They are developed to facilitate the progress of oncogenomic research. To this end, besides high-quality genomic arrays, a set of online tools are also provided for accessing and analyzing CNA data. The user friendly web interface contains search options for CNA features and patient information (Figure 9). In our implementation, raw array data with various platforms and techniques were processed into platform independent segmentation files. Importantly, the direct access to pre-computed probe level data plots supports a rapid evaluation of experiments for features of interest. Since clinical information is of importance when performing the following CNA data analysis, standardized annotation schemes were used, such as ICD classification of morphology and topography. In this way, large-scale CNA data analysis for thousands of arrays can be done efficiently and effectively.

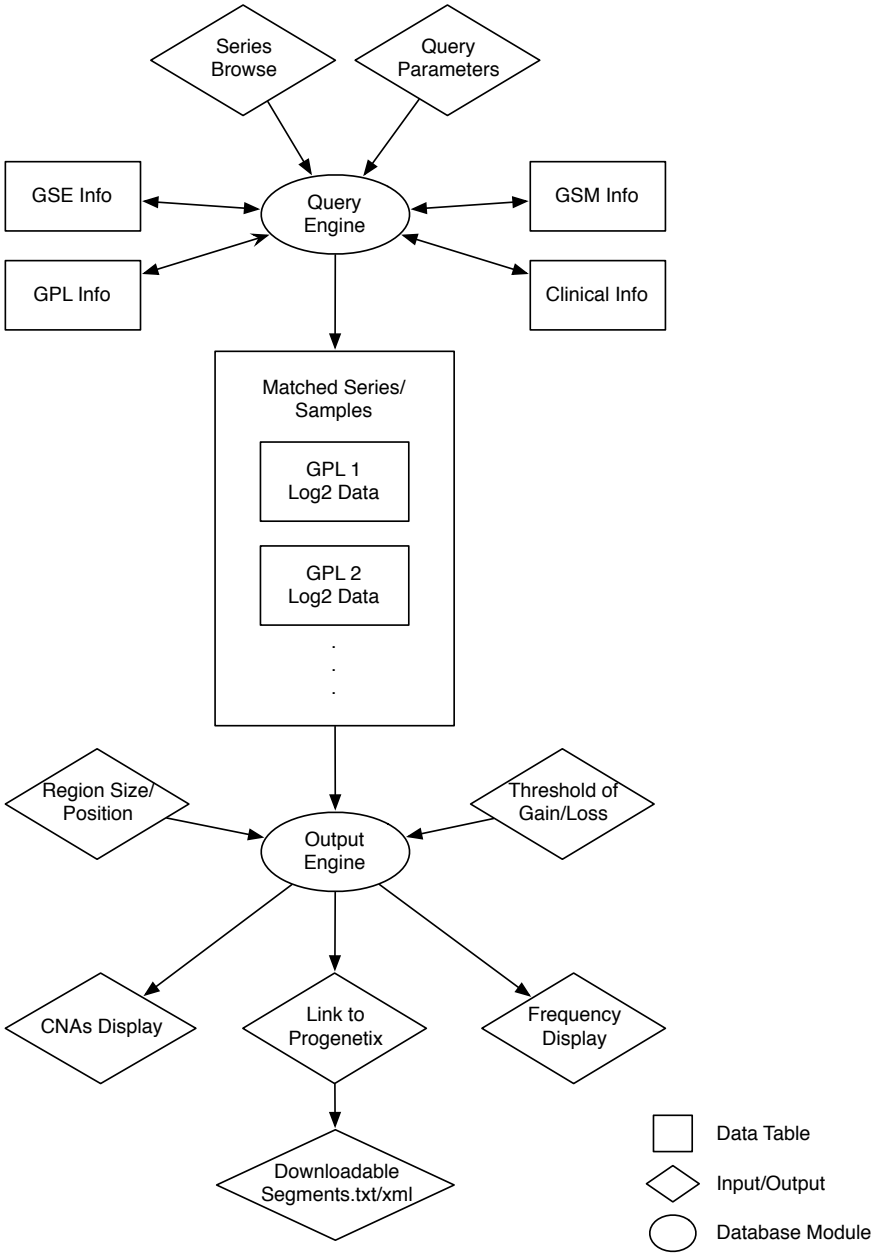


Figure 9. Overview of the basic structure of arrayMap. In the database, the two main models are query and output engines. Several types of information can be used by the query engine to search the interested CNA data. Then, matched records are combined and grouped by the output engine to perform statistical analysis, and displayed on the result page. In addition, the returned CNA data can be downloaded for further investigations.

Although both arrayMap and Progenetix provide copy number profiling data in cancer, they are not the same in many respects (Table 5). For data source, entities in arrayMap were obtained with in-house pipeline ran on the raw array data, whereas data in Progenetix were manually extracted from main text or supplementary files of publications. For techniques, Progenetix contains both chromosome and array CGH data, whereas

arrayMap focuses on array CGH data only. In terms of data visualization, arrayMap provides query and illustration options for probe-level intensity data, as well as the ability to zoom in specific region for a target gene investigation. Progenetix focuses on using statistical methods for describing and analyzing segment-level data. Note that, for the data re-analysis in arrayMap, we employed aroma.affymetrix R package[196] for processing Affymetrix arrays, and CBS (Circular Binary Segmentation)[191] algorithm for segmentation. There are several other tools available to implement these processes. Thus, the interpretation of the data can vary due to different low level (e.g. signal/background correction) and higher level (e.g. segmentation algorithms, regional or size based filtering) procedures. However, for Progenetix, we rely on the authors to ensure that the data they provided in their articles is well annotated, as usually only the segmentation data was extracted from publications. Furthermore, we are able to perform array quality assessment and filtering for arrayMap based on the probe-level data. But this process is not available for Progenetix. Finally, the data is organized in array and case-level in arrayMap and Progenetix respectively. Although the two databases are designed on different emphasis of their purpose of annotating the CNA data, we attempt to make them easy to use and serve as the valuable reference databases for the research community.

Table 5. A comparison between Progenetix and arrayMap

Features	Progenetix	arrayMap
Techniques	cCGH, aCGH	aCGH
Scope	case-level	array-level
Raw data presentation	no	yes; CEL file, log2, segmentation
Raw data re-analysis	no; supervised result	yes; re-segmentation, thresholding
Final data	annotated CN status for GP and cytogenetic regions	unsupervised CN status for GP and cytogenetic regions

Data from arrayMap and Progenetix can be used on different levels, i.e. locus centric and for entity profiling. Obviously, systematic analysis will help researchers to discover features which are indiscernible in individual studies, and thus bring new insights for understanding of disease pathology and the development of new therapeutic approaches. Since both databases support batch data download, researchers are able to integrate these data with

their own analysis efforts, e.g. to increase sample size or for result verification purposes. I believe these two databases will promote further evolution of microarray data meta-analysis. Furthermore, arrayMap includes more than 3000 normal samples from healthy individuals or from normal tissues of cancer patients. These normal control samples are of vital importance for genomic imbalances studies. They could be integrated as reference dataset, e.g. to account for copy number variation data superimposed on the tumor profiling results, or to act as the “base line” in some CNA calling algorithms.

During the raw data processing procedure, I encountered a large number of individual data sets with insufficient or limited probe quality in both GEO and ArrayExpress. This may be due to the lack of quality control step for GEO and ArrayExpress, since they aim at storing raw experimental data. These poor quality arrays may lead to erroneous analysis results, especially in certain quality sensitive study such as chromothripsis pattern identification. Currently, the representative genomic array quality control methods are NUSE (Normalized Unscaled Standard Error) and RLE (Relative Log Expression). They were developed from fitted statistical models of probe level data. NUSE provides a measure of relative array quality derived from the residuals from the Robust Multichip Analysis (RMA) model. RLE is an absolute metric that measures by summarizing the distribution of relative log expressions within a set of arrays against a reference set. They provide a better basis for judging quality compared to standard methods like SNP Call Rate. However, NUSE can only be used to assess the relative quality of arrays within a dataset, whereas the results obtained from RLE will be influenced by different batches of experiments. In general, they have a limited accuracy for detection of arrays with inadequate probe-level data. Currently, the most viable strategy for quality assessment of processed, heterogeneous copy number arrays is the visual inspection of probe plotting and segmentation results through an experienced researcher. I generated a quality classification system for arrayMap. Each array in the database was classified into four categories based on inspections of genome-wide array plots. Depending on the intended research purpose this basic classification system can be used for a pre-analysis triage of copy number data. In addition, these quality tags can also be used as the “gold standard” for further quality control algorithm design. A platform independent quality assessment system for genomic arrays is under development at the moment.

Through the data collection and annotation efforts, it is able to provide an estimation of content and trends for the platform usage. In the last two decades, the most widely

available array CGH platforms are either based on large insert clones (BAC/P1 arrays) or based on shorter single-stranded DNA molecules (oligonucleotide arrays), which may or may not include single-nucleotide polymorphism specific probe sequences (SNP arrays). Also, although designed for gene expression profiling, cDNA arrays were used by several laboratories for measuring genomic copy number changes. In reviewing the technical platform composition, two related trends were observed. Originally developed in groups with expertise in molecular cytogenetics and cancer genome analysis, printed large insert clone arrays (BAC/P1) were the first whole genome CNA screening tools with a spatial resolution surpassing that of chromosomal CGH. However, over the last years the overwhelming use of various industrially produced oligonucleotide array platforms can be observed. These SNP arrays contain a probe density at 1-3 orders of magnitude higher than common for BAC/P1 arrays. Therefore in the near future, the very high resolution CNA data may accumulate rapidly. It will become possible to integrate these data with next-generation sequencing data and/or high density DNA methylation array data by systems biology methods to elucidate effects of genomic instability, and describe the results from more perspectives. Envisioned examples would be, e.g. the identification of genes that are involved in metastasis and treatment response; identification of chromosomal breakpoints distribution in cancer; and modeling functional networks in cancer by systems biology approaches.

7.2 ORIENTATION OF GENOME INSTABILITY

In the project described in Chapter 5, I have presented that both genes and CNAs are often clustered into hotspots, and have explored the underlying correlation between gene distribution and copy number alteration profiles across cancer genomes. To achieve this, we collected more than 16,000 cancer samples from 3 public resources of microarray data sets. Notably, focal CNAs were significantly enriched in gene-rich regions. As another manifestation of genome instability, DNA breakpoints also followed this trend. It provides us a global insight into the relationship between cancer genome instability and structure from a new perspective. The enrichment reveals that there is a non-neutral selection pressure for CNA regions across the genome. Due to observed heterogeneity of CNAs in cancer genomes, individual tumor probably follows a distinct path towards tumorigenesis.

Many genes in CNAs may lose their functions or take on new roles to promote clone expansion. Genome instability could generate enough variation, on which these non-neutral selection events can operate during tumor evolution. A full understanding of how these events contribute to specific tumor will require further studies to investigate the differential expression of genes in CNA regions.

A negative correlation between arm-level CNAs and the size of arms is observed from the entire data set, consistent with previous studies[47]. To further confirm the negative correlation with tumor specific data, I looked into arm-level CNAs of 6 cancer types. The underlying mechanism for this observation is unknown. It may reveal additional negative selective pressure on gene-rich arms, or inherent low arm-level CNA ratios of these arms.

To avoid bias, I performed cancer type and platform specific analysis across the entire dataset. Generally, focal CNAs in most cancers present an enrichment in gene-rich regions, although the extent of enrichment is a little bit different. A couple of cancers, such as hematological malignancies, present an inherent low-level of copy number changes, especially of focal CNAs. Accordingly, these cancers show relatively low correlation coefficients. Moreover, our data set is generated from 180 array platforms with various resolutions. Due to increased probe numbers, high resolution platforms are more sensitive to small copy number changes, thus usually show a substantially higher number of alterations. Therefore, compared to low resolution platforms, a higher proportion of segments that derived from high resolution arrays fit our definitions of focal CNAs.

In summary, the observed significant positive correlation between genomic instability regions and gene distribution in cancer genomes may enable a better elucidation of mechanisms by which CNAs contribute to tumor development, and eventually promote a more systematic understanding of human cancer.

7.3 CHROMOSOME SHATTERING AND CELL FATE

Since the initial report of chromothripsis over two years ago, many following studies have been performed and provided several hypotheses of this phenomenon. Now it has been characterized as a type of focally clustered genomic aberration events, occurring in a small subset of cancer genomes. Our recent work described in Chapter 6 identified 918 chromothripsis-like genome profiles, based on an analysis of copy number aberration patterns from 22,347 oncogenomic arrays, representing 132 cancer types. Despite the inherent limitations of such a meta-analysis approach, this project is able to provide several new insights regarding the distribution of chromothripsis-like patterns and to produce a comprehensive estimate of chromothripsis incidence in a large range of cancer entities. This work partly overcomes the limitation of individual research resulting from the relatively small number of chromothripsis samples available.

To automatically identify chromothripsis pattern, I developed a scan-statistic based algorithm[196]. A maximum likelihood ratio score was employed, which is commonly used to detect clusters of events in time and/or space and to determine their statistical significance. According to previous studies, segmental copy number status changes and significant breakpoint clustering are two relevant features of chromothripsis. According to these features, in our implementation, the algorithm used a series of sliding windows to identify the chromosome region with the highest likelihood ratio as the chromothripsis candidate. Note that there is another algorithm in a recent published study can be used to detect chromothripsis event. In that study, authors identified chromosomes with at least 10 copy number alterations, and breakpoints have to be randomly distributed over the chromosome. They also requested the sizes of neighboring segments to be roughly the same or the same order of magnitude. Indeed, this model effectively detected chrompthripsis-like cases. However, the model is not able to distinguish the exact spot on which chromosome smashed. These pulverization regions are important for examining associations between tumor type related cancer associated genes and molecular mechanisms behind chromothripsis events. For example, in our dataset, chromothripsis occurs more frequently in chromosome 17 than in any other chromosome. This observation is in accordance with data reporting an association between chromothripsis and *TP53* mutations in Sonic-Hedgehog medulloblastoma and acute myeloid leukemia[125]. *TP53* is located on the p arm of chromosome 17, and is involved in cell cycle control, genome maintenance and apoptosis. Based on the observation, we hypothesize that *TP53*

mutation is a recurring and possibly associated event in chromothripsis formation. Therefore, the accurate identification of chromosome shattering region may point to specific driver mutations that contribute to chromothripsis events or to a class of chromothripsis derived cancer promoting mutations.

In order to test the performance of our algorithm, 23 previously published chromothripsis cases with available raw array data were collected and used as a training set. I generated a receiver operating characteristic (ROC) curve from the training set results, and selected optimal cutoff values based on this curve. Not many published studies provide raw array data to allow us to perform re-analysis. Furthermore, until now, a lot of chromothripsis related research are based on whole-genome sequencing data. These data provide additional information that can not be obtained by microarray, such as the class of rearrangements involved in the chromosome shattering region. The type of breakpoint junctions, such as nonhomologous end joining or microhomology-mediated end joining, may reveal break repair pathways, which are important for understanding mechanisms underlying chromothripsis[129]. Moreover, inter-chromosomal or intra-chromosomal breakpoint junctions are also only able to be detected by genome sequencing. According to previous research, at least two mechanisms driving formation of genomic rearrangements may contribute to the occurrence of chromothripsis: double-strand DNA breaks and template-switching during DNA replication. In our dataset, about one-fourth of chromothripsis cases affected at least two chromosomes, which, from another point of view, confirmed that more than one mechanisms may be responsible for this catastrophic event. For certain candidate mechanisms, e.g. micro-nucleus formation due to mitotic delay, this observation would imply more than one event whereas the observation appears compatible with an aborted apoptosis process.

In the initial study, the authors stated that chromothripsis could be a one-off cataclysmic event that generates multiple concurrent mutations and rearrangements[105]. Strikingly, and in possible contradiction to this proposed singular “shortcut” to cancer genome generation, we observed in the majority of chromothripsis samples additional, complex non-CTLP genome re-arrangements. Plausible and non-exclusive explanations could be that chromothripsis might frequently arise due to previously established errors in the maintenance of genomic stability, or that chromothriptic aberrations involving genomic maintenance genes may predispose to the acquisition of additional CNA. For those frequent cases exhibiting additional non-chromothripsis CNA events, their possible

contribution to oncogenesis has to be considered when modeling the role of chromothripsis in cancer development. Further efforts are needed to investigate the temporal order of chromothripsis and non-chromothripsis events observed in genomically complex samples.

At last, our dataset revealed that chromothripsis are related to overall more advanced tumor stages and overall worse prognosis, when compared to non-chromothripsis cases. One possible explanation is that the numerous concurrent genetic alterations induced by chromothripsis events disturb a larger number of genes and contribute to more aggressive tumor phenotypes. By themselves, these observations do not differentiate whether chromothripsis is an early event promoting aggressive tumor behavior with fast growth rates and reduced response rates to therapeutic interventions; or whether this observation relates to underlying primary mutations predisposing to genomic instability, aggressive clinical behavior and chromothripsis as an epiphenomenon. Interestingly, the high rate of *TP53* involvement by itself would support both possibilities for this gene, i.e. chromothripsis as result of *TP53* mutation as well as chromothriptic events with *TP53* locus involvement promoting an aggressive clinical behavior[122].

In conclusion, chromothripsis represents a striking feature occurring in a limited set of cancer genomes, and can reliably be detected using biostatistical methods. The observed patterns may reflect on heterogenous biological phenomena beyond a single class of “chromothripsis” events, and probably vary in their specific impact on oncogenesis. Fragmentation hotspots derived from our large-scale data set may promote the detection of markers involved in chromothriptic rearrangements, or may be used for assigning disease related effects to a chromothripsis induced genomic events.

8 PERSPECTIVES

8.1 EXPANSION OF CANCER GENOME PROFILES IN THE FUTURE

Thanks to the widespread implementation of array comparative genomic hybridization and SNP arrays to detect genome-wide copy number changes in cancer, the last decade have witnessed a significant increase in oncogenomic array data. For example, during the last 12 years, the size of our Progenetix database has been grown almost 60 times, with included cancer samples increasing from 490 to more than 30 000. The advantage of making use of the vast amount of genomic array data is exemplified by recent studies that performed large-scale aCGH data analyses on a wide range of cancer types, provided striking new insights into human cancers. In the future, one can imagine the continued expansion of this resource. Note that, there are thousands of cancer samples in our database with detailed clinical information and outcome of patients. Data mining of these records can provide valuable insight into the complicated process of tumorigenesis. To this end, further online tools may be developed and provided to the research community. The main purpose is to try to find aberrations that correlate with a specific diagnostic, prognostic or therapeutic trait, such as poor prognosis or drug resistance. In the meanwhile, new visualization options and components may also be added to make the website more powerful and easier to use.

In the first version of arrayMap, I generated a pipeline for determining the genomic positions for the tens to hundreds of thousands array probes with reference to a common genome Golden Path edition. For each array platform, the genome positions of probes were remapped to the current commonly used version of the human reference genome assembly (UCSC HG18). This information and procedure, presented in Chapter 4, can be updated. At the moment the latest version of the genome assembly is UCSC HG19. The accurate physical coordinates of probes are important for analyzing gene-specific copy numbers and facilitating the integration of data measured with different array platforms. In previous work, specific mapping procedures were employed for different types of probes.

In average, about 96% probes were successfully remapped (Table 6). This percentage may be improved by the up-to-date UCSC Genome annotation database.

Table 6. Percentage of remapped probes classified by platform types

Platform Type	Average Mapping Rate	Number of Arrays	Number of Platforms
Original HG18	NA	1583	40
in situ oligonucleotide	99%	21678	55
BAC/P1	98%	5464	55
spotted DNA/cDNA	91%	2365	82

So far only human cancer genome data is collected for the database. However, information of other species can be added as an extension. Although both databases are not user driven repository, several user submitted data series are already included in arrayMap. Accordingly, the website provides an interface for user data processing. Users are able to upload private data in a number of formats, and to obtain summary reports and to visualize their data. At present, the Zebrafish (Danio rerio) genome is integrated into Progenetix, which enables researchers to analyze Zebrafish data with the benefit of online tools.

8.2 “SNIPING” CANCER GENES

As discussed in Chapter 5, arm-level genomic aberrations often involve entire chromosomes or chromosomal arms. These patterns are typically tumor type specific and have been used to distinguish cancer subtypes. Since arm-level CNAs spread over thousands of genes, focal CNAs may be more useful to narrow the list of candidate targets. With the progress of microarray technology, current arrays contain as many as 1.8 million probes that tile the reference human genome. They enable the sensitive and reliable detection of very small (~10 kb) CNAs. Such high resolution arrays can readily

detect CNAs spanning several genes or even single gene. Therefore, the identification of candidate cancer genes by focal CNAs is possible. Many oncogenes and tumor suppressors have been identified from copy number profiles of cancer genomes. For example, GISTIC (Genomic Identification of Significant Targets in Cancer)[50] is an algorithm designed to identify significantly altered genomic regions. The detected peak regions frequently contain target genes. Based on our large-scale datasets, focal CNAs can be a valuable resource in identifying novel cancer genes.

Besides focal CNAs, breakpoints are also potential resources for screening driver genes. For instance, DNA strand break may occur within a gene and result in dysfunctional protein production. If breaks occur within two distinct genes to create a fusion gene, it may encode either a defective protein or a protein with novel functions. In addition, breakpoints may disturb regulatory elements of genes, so that change the expression of the involved gene. Note that, the ability to detect target genes by breakpoints will be limited by resolution of arrays, the higher resolution, the more accurate localization of breakpoints can be obtained. Moreover, CNVs may also influence the results and cause false positives. A recent study developed an algorithm, GFD (Genomic Fusion Detection)[197], to detect fusion genes based on the segmentation data from Affymetrix SNP6 array. In this tool, breakpoints are defined by segments of DNA. Subsequently, simulated and real cancer data were used to train and test this algorithm. GFD is able to detect non-functional, silenced and novel fusions. In general, many useful tools are available to be applied to our datasets. They are supposed to provide further valuable data.

8.3 DIGGING DEEPER INTO THE MECHANISMS OF CHROMOTHRIPSIS

In the work described in Chapter 6, we identified 1,269 chromothripsis chromosomes from 918 cases among more than 100 cancer types. This dataset gives us a good opportunity to further explore the mechanism(s) responsible for the generation of chromothripsis. In our analysis, although most (76%) chromothripsis cases presented single chromosome shattering events, in approximately 24% chromothripsis cases affected at least 2 chromosomes (Figure 10A). According to this observation, we hypothesize that more than one mechanism may be responsible for this catastrophic event.

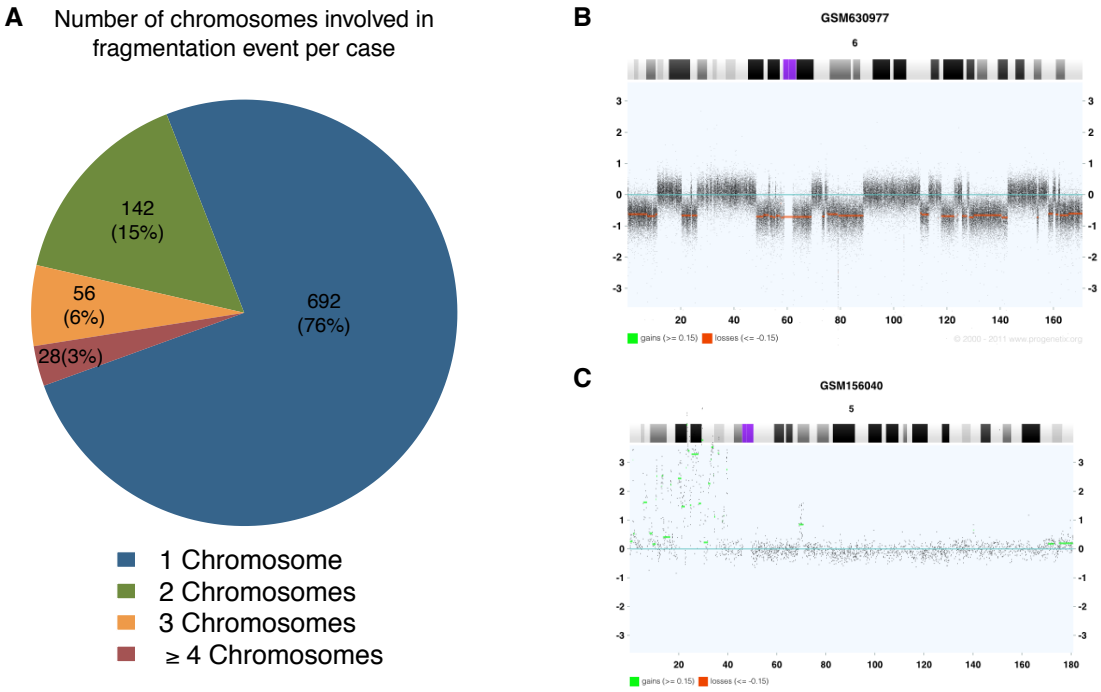


Figure 10. Evidence of different potential mechanisms underlying chromothripsis. (A) The number of chromosomes affected by chromothripsis per sample. The numbers outside and inside the brackets are number and percentage of CTLP samples respectively. (B) An example of chromothripsis that may result from DNA double-strand break. (C) An example of chromothripsis that may be caused by replicative processes.

Interestingly, the distinctive patterns of copy number states give some clues for the mechanisms underlying these events. Chromothripsis cases in our dataset can be classified into two basic categories, based on their copy number states of the pulverized genomic regions[105,198]. In the first category, the copy number changes alternate between two or three states: gain, loss and/or normal (Figure 10B). It implies that these breakpoints may result from DNA double-strand break while the chromosomes are condensed for mitosis[105,107]. This hypothesis is supported by the initial report of chromothripsis phenomenon. The second category comprises of cases which exhibit more than three copy number states in the chromosome shattered regions[35,41,144] (Figure 10C). These features may be attributed to long-distance template-switching by FoSTeS[42] (Fork Stalling and Template Switching) or MMBIR[36,39] (Microhomology-Mediated Break-Induced Replication). The evidence in support of this hypothesis comes from a recently published study that revealed the replicative processes involved in chromothripsis by breakpoint sequencing. To further elucidate these potential mechanisms, the cancer-related genes

and pathways in chromothripsis regions can be investigated. They may provide insight into the difference between both mechanisms. Furthermore, the measurement of the enrichment with respect to GO (Gene Ontology)[199] categories can be performed to facilitate biological interpretation and hypothesis testing. Since the chromosome-shattering was observed in a wide variety of tumors, the underlying mechanism is likely to reflect unknown general features of human cancer.

Part IV

APPENDIX

Specific Genomic Regions Are Differentially Affected by Copy Number Alterations across Distinct Cancer Types, in Aggregated Cytogenetic Data

OPEN ACCESS Freely available online

PLOS ONE

Specific Genomic Regions Are Differentially Affected by Copy Number Alterations across Distinct Cancer Types, in Aggregated Cytogenetic Data

Nitin Kumar[1,2◑], Haoyang Cai[1,2◑], Christian von Mering[1,2]*, Michael Baudis[1,2]*

1 Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland, **2** Swiss Institute of Bioinformatics, Quartier Sorge, Lausanne, Switzerland

Abstract

Background: Regional genomic copy number alterations (CNA) are observed in the vast majority of cancers. Besides specifically targeting well-known, canonical oncogenes, CNAs may also play more subtle roles in terms of modulating genetic potential and broad gene expression patterns of developing tumors. Any significant differences in the overall CNA patterns between different cancer types may thus point towards specific biological mechanisms acting in those cancers. In addition, differences among CNA profiles may prove valuable for cancer classifications beyond existing annotation systems.

Principal Findings: We have analyzed molecular-cytogenetic data from 25579 tumors samples, which were classified into 160 cancer types according to the International Classification of Disease (ICD) coding system. When correcting for differences in the overall CNA frequencies between cancer types, related cancers were often found to cluster together according to similarities in their CNA profiles. Based on a randomization approach, distance measures from the cluster dendrograms were used to identify those specific genomic regions that contributed significantly to this signal. This approach identified 43 non-neutral genomic regions whose propensity for the occurrence of copy number alterations varied with the type of cancer at hand. Only a subset of these identified loci overlapped with previously implied, highly recurrent (hot-spot) cytogenetic imbalance regions.

Conclusions: Thus, for many genomic regions, a simple null-hypothesis of independence between cancer type and relative copy number alteration frequency can be rejected. Since a subset of these regions display relatively low overall CNA frequencies, they may point towards second-tier genomic targets that are adaptively relevant but not necessarily essential for cancer development.

Citation: Kumar N, Cai H, von Mering C, Baudis M (2012) Specific Genomic Regions Are Differentially Affected by Copy Number Alterations across Distinct Cancer Types, in Aggregated Cytogenetic Data. PLoS ONE 7(8): e43689. doi:10.1371/journal.pone.0043689

Editor: Patrick Tan, Duke-National University of Singapore Graduate Medical School, Singapore

Received April 30, 2012; **Accepted** July 23, 2012; **Published** August 24, 2012

Funding: The authors have no funding or support to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mbaudis@imls.uzh.ch (MB); mering@imls.uzh.ch (CvM)

◑ These authors contributed equally to this work.

Introduction

Genetic changes such as point mutations, regional copy number alterations/aberrations (CNA) and structural changes (e.g. gene fusion events) are all hallmarks of cancer. CNAs arise as somatic changes in the tumor cell genome through a variety of mechanisms and can be observed in virtually all types of cancer, to a varying extent. So far, the most widely used methods for the detection of CNAs have been chromosomal and array-based Comparative Genomic Hybridization (CGH) techniques [1–4]. Localized, recurring CNAs (hot-spots) have been shown to target canonical oncogenes (e.g. duplications/amplifications of the MYC, MYCN, REL loci) or tumor suppressor genes (e.g. deletions of the CDKN2A/B, TP53, ATM loci). Some regional CNAs such as gains on 8q and losses on 3p are present across multiple cancer types, whereas other imbalances may be largely restricted to a limited number of cancer entities [5].

Datasets integrated across multiple cancer types have previously been analyzed, to report regional "hot-spots" of frequent CNAs [5,6]. In a given set of individual tumor samples, the number and distribution of CNAs varies considerably [5] and this genetic heterogeneity has been used to detect and report co-occurring CNAs [7].

In principle, specific patterns and similarities in the individual and/or disease specific CNA profiles might point to distinct oncogenomic mechanisms acting in different cancer types and specimens, given a sufficiently large number of data points. Indeed, clustering of CNA patterns has been used to identify oncogenomic similarities [5,8–11]. The adaptation of clustering techniques to the analysis of CNA patterns has been subject of previous studies [12–14]. With a few exceptions [5,14], however, sample-based clustering has been the main focus of such studies so far. In contrast, we here explore the clustering of cancer types, not of individual cancer samples.

Both descriptive and clustering-based analyses of CNA across multiple cancer types suffer from a bias towards the more frequently occurring events. Due to the heterogeneity of the overall CNA signal, with greatly varying average frequencies of

CNAs per cancer type (Figure 1a), clustering results may be distorted depending on the disease entities analyzed. This variation in overall CNA occurrence frequencies across cancer types may simply be owed to differences in the average time points of clinical detection or in different progression characteristics, and should be corrected for prior to clustering analyses. To the best of our knowledge, so far no implementation has been reported for a comprehensive, very large-scale clustering analysis of frequency-normalized cancer CNA profiles.

Here, we focus on the identification of genomic regions that contribute meaningfully to the clustering of cancer types. From hereon we will refer to those as "non-neutral" regions. As the starting point of our analysis, we use hierarchical clustering to arrange cancer types on the basis of their CNA frequency profiles. We then employ a permutation approach to estimate the relative contribution of individual genomic regions to the quality of the clustering and to the derived relationship tree. The clustering quality is inferred from an intrinsic measure (summed branch lengths: tree height statistics), and genomic regions that reject the null hypothesis are termed non-neutral. Identified regions are compared to canonical CNA hot-spots (i.e. those that occur most frequently across the entire dataset).

Our current analysis is based on data from a total of 25579 samples, which are classified into 160 different cancer entities (table S1) according to the International Classification of Disease in Oncology (ICD-O 3). Our approach is unique in that it a) focuses less on the clustering as such but more on the individual genomic regions that best support the clustering, b) uses an intrinsic quality measure coupled to a permutation strategy for validation, c) performs CNA frequency normalization prior to analysis, and d) is based on a very large data set, processed in a standardized setup. We aim for the identification of potential cancer-specific driver/modulator regions, which may not have been detected in earlier, largely hot-spot-focused approaches. All of the underlying cancer data is available through our Progenetix repository (www.progenetix.org; [15]).

Results

The average overall frequency of CNAs across the entire genome varies among different cancer types (Figure 1a). Since the relative weight of CNAs at individual genomic regions in a given cancer type depends on the observed overall genome-wide frequency, we aggregated all patient samples by cancer type and normalized the frequencies of CNAs for each cancer type to the overall mean observed across the entire data set (Figure 1b, Figure S1). The normalized CNA frequency profiles were then clustered using hierarchical clustering.

To evaluate the quality and the biological signal in the clustering, we labeled each cancer type with its "root" cell type (i.e., an undifferentiated cell type from which the tumor likely originated). We expected cancers of the same root cell type to cluster together; this was used as an external proxy for the expected biological relationships between cancer entities. The Random Index [16] was used to compute this external cluster quality measure. Tumors of the same cell type indeed often clustered together, usually in 2–3 small groups (Figure 2). The consistency of this grouping was significantly higher than expected

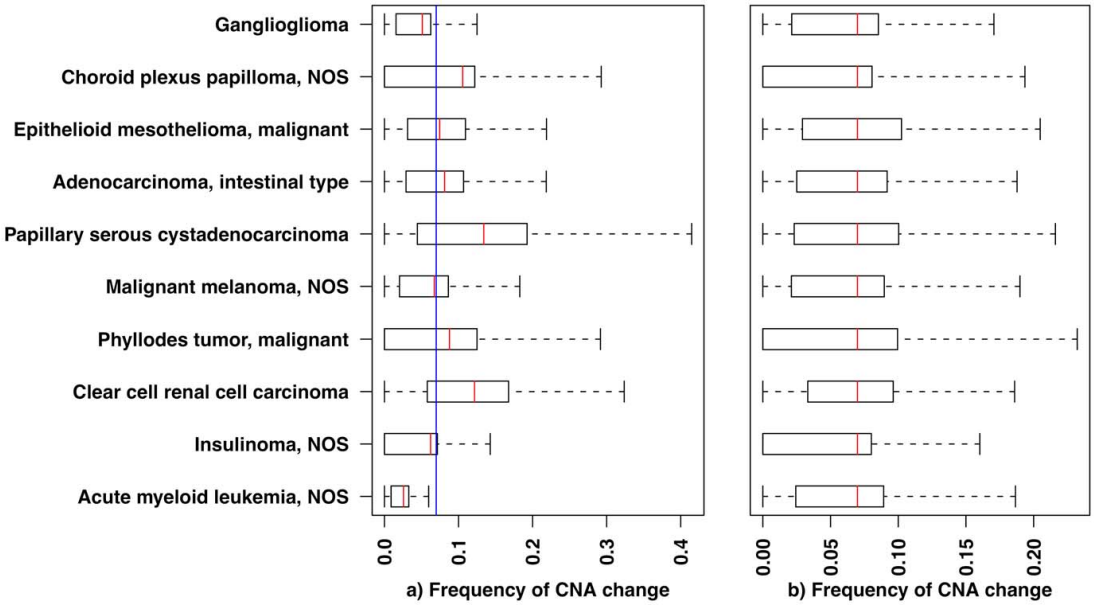


Figure 1. The overall frequency of genomic copy number alterations (CNA) differs among cancer types. Boxplots show the CNA frequency distributions among tumor samples in 10 randomly selected cancer types. The boxplot delineations mark the percentiles 5%, 25%, 75% and 95%. The red lines indicate the mean frequency for each cancer type, whereas the blue line represents the overall mean frequency across all 160 cancer types analyzed here. Frequency values are defined as the ratio of number of samples showing a CNA for a genomic region (i.e., cytogenetic bands) over total number of samples in that cancer type. a) Before normalization b) After normalization. In b) the nominal frequency distribution for each cancer type is re-scaled so that its mean matches the overall mean across all cancer types. (NOS – "not otherwise specified": high-order classifications, not further assigned to more detailed levels).
doi:10.1371/journal.pone.0043689.g001

at random, pointing towards biologically meaningful differences in CNA profiles between tumors of distinct origins. Cutting the tree at several heights always led to an observed quality of clustering that was better than the expected random value (Figure 2), except for the cut at the highest level, which resulted in only three clusters. This strongly argues against a completely neutral occurrence pattern of CNAs in the genome, and supports a correlation between biologically meaningful groups of cancer entities and their CNA profiles.

Randomizations of the entire frequency matrix lead to a complete loss of the signal present in the clustering tree (Figure S2), and also strongly reduced the summed branch lengths tree-height statistic.

Non-neutral CNAs

The normalized and clustered frequency matrix encompassing 160 large-scale genomic regions and 160 cancer types is shown in Figure 3. To determine how much each individual genomic region contributes to the overall signal, we individually randomized its profile across cancer types, while keeping the rest of the data unchanged. We then examined the concomitant reduction in the tree length statistics (TLS) of the clustering dendrogram, upon 100000 independent randomizations, to determine the statistical significance of that region's contribution. The resulting cancer-diverging CNA regions are important as they cannot be fully neutral and have the potential to define relationships among cancer types. Indeed, 43 out of the 160 genomic regions (table S1) were observed to have a non-neutral contribution (Bonferroni-corrected p-value ≤ 0.016) in the aggregated cancer CNA data. Note that gain and loss events were treated independently, and no preferential bias towards gains or losses was observed among the detected non-neutral regions (22 gains and 21 losses). The CNA occurrence frequencies of the non-neutral genomic regions spread thorough the entire frequency spectrum (Figure 4). Only 13 (8 gains and 5 losses) of the non-neutral regions were found altered overall more often than average (Figure 5, intersection of black and grey rectangle), indicating that subset of frequently altered hotspot regions carry a detectable signal to distinguish cancer types (the number of frequently altered regions stands at 59; Bonferroni-corrected p-value ≤ 0.016, table S1). This observation emphasizes our key point that not only the frequent CNA regions should be used to cluster and annotate cancer types.

22 genomic intervals across 12 chromosomes were found to be informative when specifically considering duplications/gains only (Table 1 and Figure 5). All three genomic segments of chromosome 18 (18p1, 18p2, 18q2) exhibited a signal. For other chromosomes such as chromosome 1 (1q2,1q3,1q4,1p2), chromosome 3 (3q1, 3q2, 3p1), chromosome 12 (12q1,12q2) and chromosome 21 (21p1, 21q1) more than 50% of genomic regions were informative as gains, suggesting simultaneous involvement of multiple loci from these chromosomes. Changes on chromosome 1(1p2), chromosome 3 (3p1, 3q1), chromosome 5 (5q2, 5q3), chromosome 9 (9p1), chromosome 11 (11p1), chromosome 12 (12q1, 12q2), chromosome 18 (18p1, 18q1, 18q2) and chromosome 21 (21p1, 21q1) were selectively informative only as gains. In terms of deletions/losses, 10 chromosomes encompassing 21 genomic regions were found to be non-neutral. Like for chromosome 18 gains, the complete chromosome 7 (7p1, 7p2, 7q1, 7q2, 7q3) was found to be informative when lost (Table 1). Informative regions on chromosome 1 (1p1,1q1, 1q2, 1q3, 1q4) and chromosome 9 (9q1, 9q3, 9p2) covered more than 50% of genomic segments present on these chromosomes. Selective losses were observed on chromosome 1 (1p1, 1q1), chromosome 6 (6q2), 7 (7q1, 7q2, 7q3, 7p2), 8 (8q1, 8q2), 9 (9p2, 9q1, 9q3), 12 (12p1), 16 (16q1). CNAs involving chromosome 1 (1q2, 1q3, 1q4), chromosome 3 (3q2), chromosome 7 (7p1), chromosome 19 (19p1) and chromosome 22 (22q1) were informative both as gain and loss events. This represents a small proportion (16%) of non-neutral CNA. Involvement of a region both as gain and loss may point towards multiple adaptively relevant loci, and/or towards a generally unstable nature of these regions.

Cancer Diverging Nature of Non-neutral CNA

To provide few examples of cancer classifying behavior of non-neutral changes, we selected a few of the enriched changes and analyzed them for their specific occurrence in different cancers. An example include cancer entities showing predominant losses versus gains on 7q. Preferential losses involving 7q were observed in germ cell, myeloid and myeloproliferative tumors (Figure 3) whereas neuroepithelial brain tumors (among other entities) preferentially displayed gains on 7q. Losses involving 7q are common in myeloid and myeloproliferative tumors [17–20] and are associated with advanced age and resistance to therapies [21,22]. However, here we show that 7q losses are quite specific to myeloid tumors and promote their selective divergence from other cancer types. 7q losses in germ cell tumors had not been explored in detail [23,24]. With the accumulation of 7q losses virtually restricted to myeloid/myeloproliferative neoplasias and germ cell tumors and in contrast to chromosome 7(q) gains observed in e.g. neuroepithelial brain tumors, it is tempting to propose involvement of at least one common oncogenetic mechanism acting in these clinically unrelated malignancies.

Chromosome 8q gains can be observed in the majority of cancer entities [5,6]. However, in our analysis 8q losses were enriched as non-neutral events. Preferential losses involving 8q were present in some brain tumors (e.g. medulloblastoma, Figure 3), separating them from other epithelial tumors. Differences in preferential losses involving 8q separated neuroepithelial tumors in two categories with both having gains on 7q but only one (mainly meduloblastomas) having preferential losses on 8q (Figure S3). Losses involving chromosome 8q across medulloblastomas have been reported by a few [25] studies before. Our analysis shows that 8q losses are selected for in some medulloblastomas and therefore could be important for cancer development/progression. Preferential losses of 8q were also observed in germ cell tumors separating them from other epithelial neoplasias (Figure S4).

As another example of restricted CNA types we also looked for cancers showing gains involving chromosome 18. Follicular lymphomas exhibited specific gains on chromosome 18 where as epithelial tumors prefered to loose chromosome 18 (Figure S4). Chromosome 18 gains are very common in follicular lymphomas and are supposed to provide an alternative mechanism for BCL2 activation [26,27]. However, here we show that this CNA event statistically separates them from other cancer types.

Discussion

Our current study represents the largest analysis performed to date on cancer CNA data, with the aim of detecting oncogenomic features that may be specifically associated or enriched in certain subsets of cancer entities. In contrast to gene-centric approaches, our analysis assesses the complete information space of genomic copy number imbalances from whole genome profiling experiments.

Overall, the frequency of CNAs across genomic intervals varied between between 0.01% to 23% (Figure 4). Clustering of cancer types on the basis of their frequency profiles helped to identify a

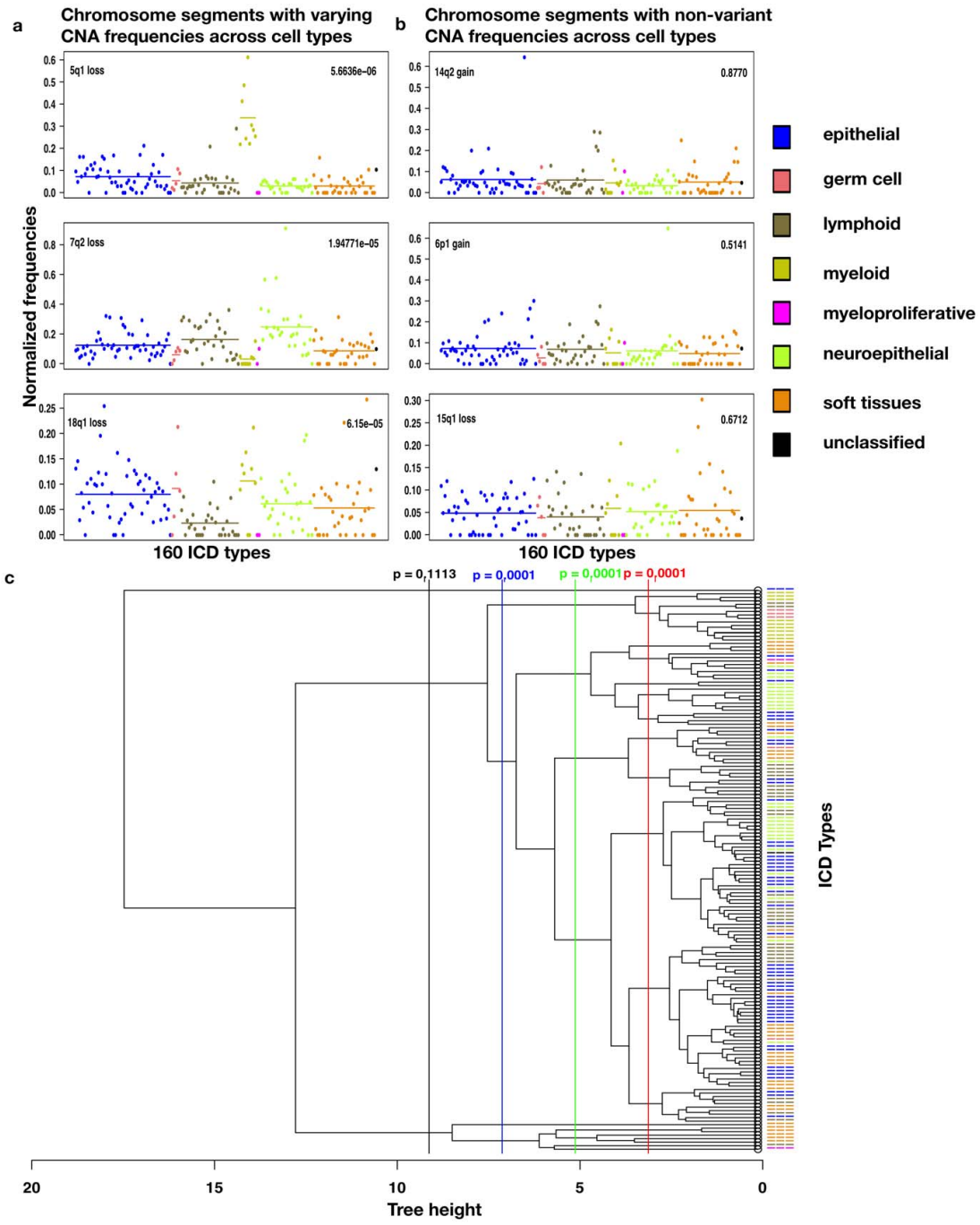


Figure 2. The tissue type of a cancer has a strong influence on its CNA likelihood pattern. a) examples of individual chromosome segments, showing their observed CNA frequencies stratified by cell type. Each dot summarizes all samples classified under one particular ICD type, color-coded by root cell type. In the left panel, three chromosome segments are shown that exhibit strong differences between cell types; on the right, three negative examples without such a signal. All p-values were corrected for multiple testing according to Benjamini-Hochberg. b) the

dendrogram (tree) has been obtained using hierarchical Ward clustering on the global frequency-normalized CNA profiles across all 160 genomic regions. Cancer types are again color-coded according to the cell type of origin, with the same legend as in a). Partitioning the tree by cutting at different heights produces multiple clusters; validation of those clusters based on the cancer origin (metric: Random Index) shows that the clustering works significantly better than expected at random.
doi:10.1371/journal.pone.0043689.g002

class of underlying molecular signals that is orthogonal to histological classifications or clinical categories (the latter are predominantly driven by the affected organ/tissue). Cancer types vary from each other in their CNA abundance, CNA size spectrum and degree of genomic instability. With respect to genomic coverage, large CNAs are generally frequent in cancer [6] and should not be excluded from statistical analyses of cancer genome patterns. While comparing CNA profiles of cancer types, their complexity and variation in frequencies have to be considered. When correcting for these parameters, regional CNAs defining the divergence of the overall profiles can be delineated.

We performed an analysis of a global cancer CNA dataset, identifying 43 genomic regions on 15 chromosomes as significant for CNA profile divergence in cancer types. Obviously, these changes do not cover the entire spectrum of CNA events in cancer, but define a subset of genomic regions that may have a possibly adaptive link to the distinct biology of various cancer types. These regions overlap rather poorly with hot-spot regions observed in many cancers. This suggests that hot-spot regions, though frequently associated with canonical oncogenes, may not always be very useful in aiding data-driven evaluation of cancer (sub-) types.

Disease specific studies have the potential to detect a representative spectrum of oncogenomic aberrations in the given entities. It can be expected that the cancer type specific regions highlighted with our approach had been discussed in the context of the respective publications. However, with our current study, we aim to provide a new, generalized approach at identifying genomic elements relevant in the genesis of individual cancer entities. Although here showcasing a "global" approach without entity pre-selection, our methodology may prove valuable when targeting relevant genomic separators in limited, biologically related entity sets.

Since the current analysis is based primarily on molecular-cytogenetic data from chromosomal CGH experiments with a spatial resolution of several megabases, only inferred information about the causal genes present in the non-neutral regions could be obtained. With upcoming high-resolution genomic array and/or sequencing data, similar analyses will more specifically define the non-neutral CNAs and can be valuable starting points for an integration of the results with functional pathway frameworks. We have recently announced the creation and public availability of a reference resource for oncogenomic array data (www.arraymap.org [28]), which will serve as starting point for such approaches

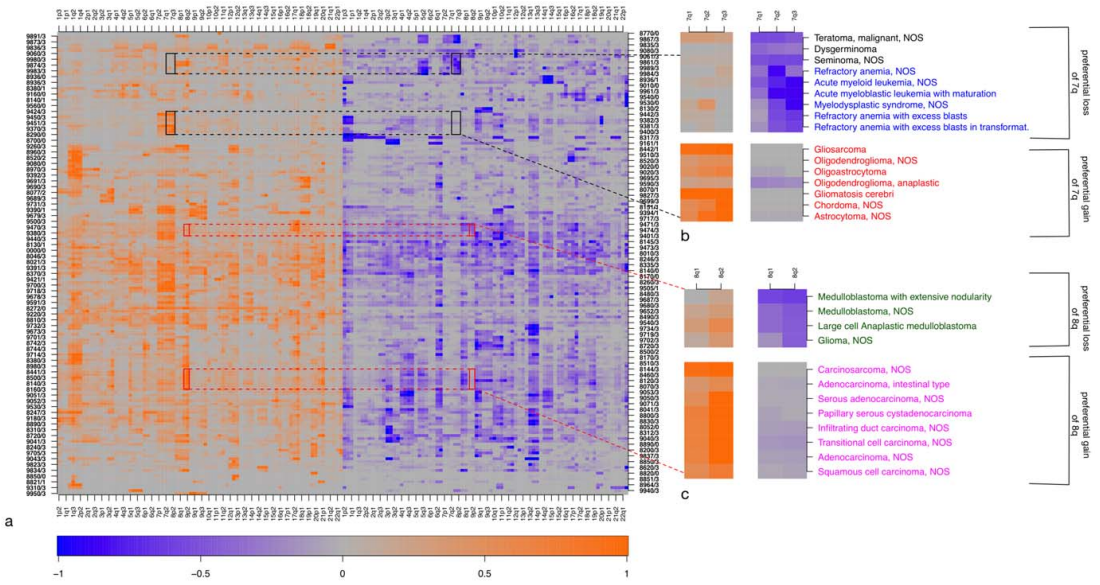


Figure 3. Examples for non-neutral CNA regions. a) Heatmap of CNA profiles on genomic regions (same clustering as in Figure 2). Genomic locations are represented with orange color when considering duplications/gains, and in blue when considering deletions/losses. Color intensity shows relative CNA frequencies; the most-affected region in each row is arbitrarily set the to brightest color (1.0) for display purposes. b) Small regions (black rectangles on the heatmap) are zoomed in to show how non-neutral CNAs can differentiate between cancer types. The example shows that 7q is preferentially gained in brain tumors (red labels) whereas it is preferentially lost in germ cell (black labels), myeloid and myeloproliferative cancer types (blue labels). c) Small regions (red rectangles on the heatmap) are zoomed in to show how 8q is preferentially lost in medullublastomas (green labels) and is preferentially gained in epithelial tumors (pink labels). Some chromosomes consist entirely of non-neutral regions (such as chromosomes 18 and 7). Note that the spatial resolution of the CNA data on the chromosome is limited (roughly corresponding to cytogenetic band resolution).
doi:10.1371/journal.pone.0043689.g003

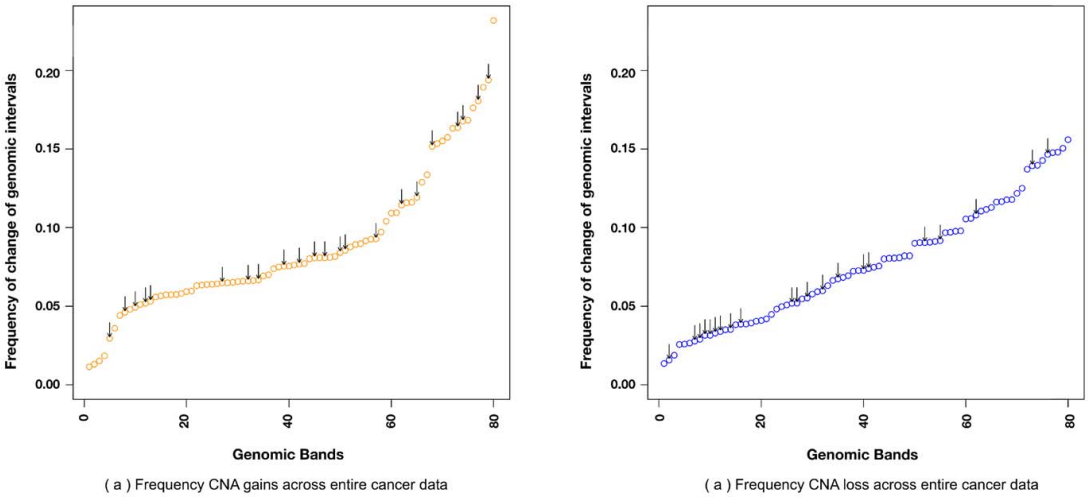


Figure 4. Not only CNA ''hotspots'' are informative in cancer classification. Genomic regions (bands) are sorted according to their overall frequency of CNAs observed. Those regions that are informative with respect to cancer type clustering are marked with arrows. a) Considering duplications (gains) b) Considering deletions (losses).
doi:10.1371/journal.pone.0043689.g004

both from our side as well as from interested members of the research community. Also, although we have focused our current analysis solely on a CNA dataset, our methodology should prove particularly valuable when combined with other sets of related diagnostics (for example point mutation data), whereby the assignment of possible driver genes in the non-neutral regions might become feasible.

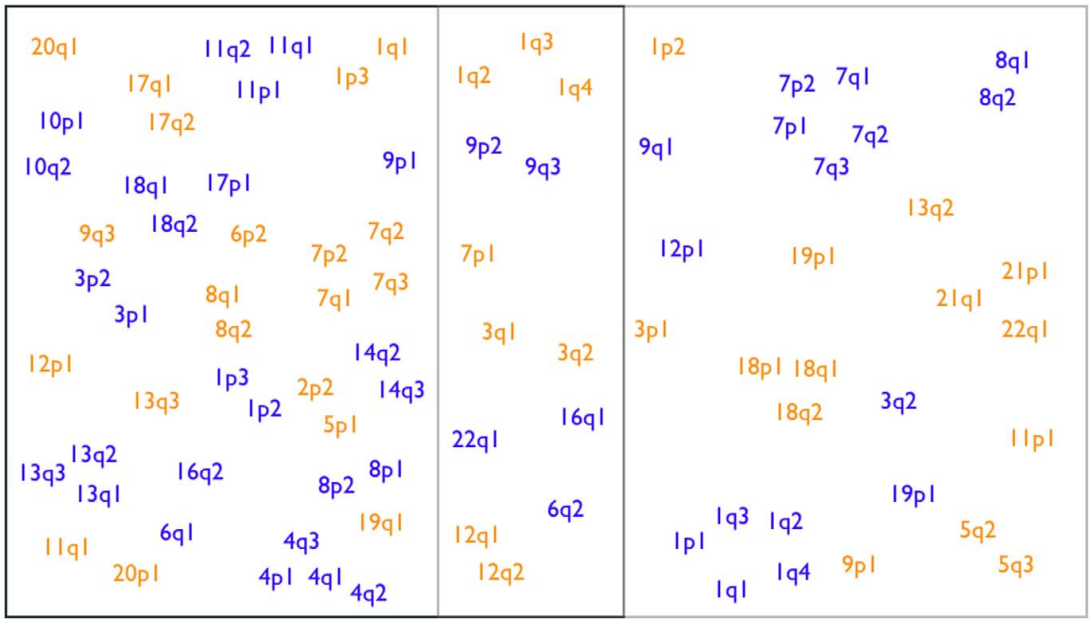


Figure 5. Comparison of non-neutral vs. hot-spot CNA. Genomic regions affected by CNAs, either more frequently than average (black rectangle), or non-neutrally with respect to cancer-type classifications (grey rectangle). The intersection defines regions that are affected both frequently and non-neutrally. Changes are color-coded (gains in orange and losses in blue).
doi:10.1371/journal.pone.0043689.g005

Table 1. Number of non-neutral regions per chromosome.

Chromosome No.	No. genomic locations	Non-neutral gains	Non-neutral losses
1	7	4	5
2	5	–	–
3	4	3	1
4	4	–	–
5	4	2	–
6	4	–	1
7	5	1	5
8	4	–	2
9	5	1	3
10	3	–	–
11	3	1	–
12	3	2	1
13	4	1	–
14	4	–	–
15	3	–	–
16	3	–	1
17	3	–	–
18	3	3	–
19	2	1	–
20	2	–	–
21	3	2	–
22	2	1	–

Some chromosomes consist entirely of non-neutral regions (such as choromosomes 18 and 7). Note that the spatial resolution of the CNA data on the chromosome is limited (it roughly corresponds to the cytogenetic banding patterns).
doi:10.1371/journal.pone.0043689.t001

Materials and Methods

Data

Our study is based on well annotated cancer CNA data from the Progenetix project [5], including a total of 25579 samples analyzed by chromosomal (cCGH; 18708) and array CGH (aCGH; 6871) experiments. The clinical samples had been classified into 160 distinct cancer entities according to the International Classification of Disease codes (ICD). At the time of writing, the Progenetix collection represents the largest resource for annotated, whole genome CNA profiling data in cancer.

For our analysis, regional CNA information across all cancer types was reduced to 80 genomic intervals covering the entire genome with the exception of the sex chromosomes. Gain and loss events were considered separately for the analysis, resulting in a matrix of dimensions nXm, where n is the number of samples and m is the number of genomic intervals (*i.e.* 160).

Cancer Clustering

The frequency of CNA changes across all genomic intervals was computed for each ICD type, and the entire frequency matrix was then normalized (Figure S1). The frequency matrix was ordered using hierarchical Ward clustering. The aggregated separation distance between cancer entities obtained using hierarchical clustering can be analyzed by parsing the clustering tree (dendrogram). The tree represents the relatedness among groups present in the same clade (similar to phylogenetic trees). Randomized data disrupts the tree completely (Figure S2), and the overall tree height statistic is reduced 3-fold, reflecting the complete loss of ordering information present in the original tree.

Method to Compare Tree Height

We used the tree height as an intrinsic measure to compare cancer associations obtained using clustering and to gauge the information present in the tree; this was used to define non-neutral CNAs. This has advantages over traditional clustering evaluation techniques, as it a) does not require external gold standard information, and b) does not require cutting the tree at an arbitrary distance. The overall tree height is defined as the sum of all direct parent-child relation path lengths in the tree. Tree distances (branch lengths) generally reflect the CNA profile discrepancies between two cancers (or groups of cancers). For any node i, the tree height between this node and its immediate parent j can be measured as $TH_j - TH_i$. The overall tree height of a tree with n nodes is than obtained as $OTH = \sum_{i=1,j=1}^{i=n,j=n} TH_j - TH_i$ (figure S3).

Tree length statistics (TLS). To identify genomic regions that are non-neutrally affected by CNA we have developed the following permutation strategy:

1. Normalized frequencies of CNA across all genomic intervals are computed across all cancer types.
2. The cancer classification tree is obtained using hierarchical Ward clustering.
3. The observed over all tree height (OTH_o) is calculated as mentioned above (Figure S5).

4. A counter C is set to zero for every genomic interval in consideration.
5. For any genomic interval i, its status values are shuffled among all samples keeping its over all frequency the same (n_i).
6. The frequency of CNA at genomic interval i is re-calculated after randomization across all cancer types. The shuffling in the previous step changes the frequency of interval i across all cancer types keeping the normalized frequency distribution of all other genomic intervals.
7. The frequencies for interval i in the normalized frequency matrix from step one are replaced with permuted frequencies for this interval and the permuted overall tree heigh ($OTH_{i,p}$) is computed.
8. If $OTH_{i,p} \geq OTH_o$, C is incremented as C = C+1.
9. p-value for genomic location i, at the end of N (100'000) permutations is computed as $p_i = C/N$.
10. p-values across all bands are corrected for false discovery rate using Bonferroni correction.

Frequency Based Enrichment (FBE)

Frequently observed CNA regions ("hot-spots") are genomic changes that occur more often than expected under a fully random null model. Such hot-spot CNAs can be identified using the binomial probability function [29]. Let's suppose genomic interval i shows a CNA across n_i samples out of N samples. The background CNA frequency (n_b) can be represented as the mean frequency change across all intervals. The p value that the frequency of CNA n_i, is more than any frequency x ($n_i \geq x$) is obtained using the binomial probability function.

$$p(n_i|N,n_b) = \binom{N}{n_i} n_b^{n_i}(1-n_b)^{N-n_i}$$

$$p_i = \sum_{n=x}^{N} p(n_i|N,n_b)$$

Genomic intervals showing a large deviation from the mean will be assigned low p-values. All p-values are corrected for false discovery rate using Bonferroni correction.

Supporting Information

Figure S1 Method for CNA frequency normalization across cancer types. All the frequencies among cancer types were normalized to the mean frequency of CAN changes across across the 160 cancer types. This normalization was achieved by multiplying the cancer-type-specific frequencies with an index A_n, whose value was calculated as shown.
(PNG)

Figure S2 Dendrogram of a permuted frequency matrix. For this clustering, the frequencies among cancer types were permuted and then normalized. Hierarchial Ward clustering was then performed and the dendrogram tree shown was obtained. The tree height is severely affected by the permutation. In this randomized clustering, similar cancer types no longer clustered together.
(PDF)

Figure S3 Small regions from heatmap in main Figure 3 are shown here. These regions represent gains and losses on 7q and 8q. 8q changes differentiate between two categories of brain tumors, with a subset showing preferential losses on 8q (green labels) and other rarely showing involvement of 8q locus (red label). Thus depending on 8q involvement neuroepithelial tumors can be divided in to two different categories. Both of them show 7q gains.
(PDF)

Figure S4 Examples for non-neutral CNA regions. a) Heatmap of CNA profiles on genomic regions (same as in Figure 3). b) Small regions (red rectangles on the heatmap) are zoomed in to show how 8q is preferentially lost in in germ cell (black labels) tumors and is preferentially gained in epithelial cancer types (pink labels). c) Small regions (black rectangles on the heatmap) are zoomed in to show how 18q is preferentially gained in medullublastomas (brown labels) and is preferentially lost in epithelial tumors (pink labels). The examples here show that how two different non-neutral changes differential epithelial tumors from germ cell tumors and follicular lymphomas.
(PDF)

Figure S5 Calculation of over all tree height. Schematic representation of the summed branch-length tree height statistic. Overall tree height is computed by summing up the distance between all parents and child nodes. Note that the branch lengths of terminal branches ("leafs") are not considered. Overall tree height = $H_{A-C} + H_{B_D} + H_{AB} + H_E$.
(PDF)

Table S1 Table with information about cancer types used in the analysis, non-neutral and hot-spot p values. The table giving details about all cancer types used in this analysis with the corresponding numbers of samples in them and the root cell type of each cancer. The table also has information about the non-neutral and hot-spot p-values obtained for all genomic bands in analysis.
(ODS)

Author Contributions

Conceived and designed the experiments: NK HC CvM MB. Performed the experiments: NK HC MB. Analyzed the data: NK HC CvM MB. Contributed reagents/materials/analysis tools: NK HC MB. Wrote the paper: NK CvM MB.

References

1. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258: 818–21.
2. Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, et al. (1993) Detection of amplified dna sequences by reverse chromosome painting using genomic tumor dna as probe. Hum Genet 90: 584–9.
3. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, et al. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 20: 399–407.
4. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20: 207–11.
5. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative metaanalysis of chromosomal cgh data. BMC Cancer 7: 226.
6. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905.
7. Kumar N, Rehrauer H, Cai H, Baudis M (2011) Cdcoca: a statistical method to define complexity dependence of co-occuring chromosomal aberrations. BMC Med Genomics 4: 21.

8. Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, et al. (2006) Dna copy number amplification profiling of human neoplasms. Oncogene 25: 7324–7332.
9. Liu J, Ranka S, Kahveci T (2007) Markers improve clustering of cgh data. Bioinformatics 23: 450–7.
10. Ferreira BI, Garcia JF, Suela J, Mollejo M, Camacho FI, et al. (2008) Comparative genome profiling across subtypes of low-grade b-cell lymphoma identifies type-specific and common aberrations that target genes with a role in b-cell neoplasia. Haematologica 93: 670–679.
11. Takeuchi I, Tagawa H, Tsujikawa A, Nakagawa M, Katayama-Suguro M, et al. (2009) The potential of copy number gains and losses, detected by array-based comparative genomic hybridization, for computational differential diagnosis of b-cell lymphomas and genetic regions involved in lymphomagenesis. Haematologica 94: 61–69.
12. Liu J, Ranka S, Kahveci T (2006) Markers improve clustering of cgh data. Bioinformatics 23: 450–457.
13. Wieringen WNV, Wiel MAVD, Ylstra B (2008) Weighted clustering of called array cgh data. Biostatistics 9: 484–500.
14. Liu J, Bandyopadhyay N, Ranka S, Baudis M, Kahveci T (2009) Inferring progression models for cgh data. Bioinformatics 25: 2208–15.
15. Baudis M, Cleary ML (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 17: 1228–9.
16. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Bosotn, MA, USA: Addison Wesley.
17. Kühn MWM, Radtke I, Bullinger L, Goorha S, Cheng J, et al. (2012) High-resolution genomic profiling of adult and pediatric core-binding-factor acute myeloid leukemia reveals new recurrent genomic alterations. Blood.
18. Woo KS, Kim KE, Kim KH, Kim SH, Park JI, et al. (2009) Deletions of chromosome arms 7p and 7q in adult acute myeloid leukemia: a marker chromosome confirmed by array comparative genomic hybridization. Cancer Genet Cytogenet 194: 71–4.
19. Cordoba I, González-Porras JR, Nomdedeu B, Luño E, de Paz R, et al. (2012) Better prognosis for patients with del(7q) than for patients with monosomy 7 in myelodysplastic syndrome. Cancer 118: 127–133.
20. Aktas D, Tuncbilek E (2006) Myelodysplastic syndrome associated with monosomy 7 in childhood: a retrospective study. Cancer Genet Cytogenet 171: 72–5.
21. Appelbaum FR, Gundacker H, Head DR, Slovak ML, Willman CL, et al. (2006) Age and acute myeloid leukemia. Blood 107: 3481–5.
22. Wong JCY, Zhang Y, Lieuw KH, Tran MT, Forgo E, et al. (2010) Use of chromosome engineering to model a segmental deletion of chromosome band 7q22 found in myeloid malignancies. Blood 115: 4524–32.
23. McIntyre A, Summersgill B, Lu YJ, Missiaglia E, Kitazawa S, et al. (2007) Genomic copy number and expression patterns in testicular germ cell tumours. Br J Cancer 97: 1707–12.
24. Veltman I, Veltman J, Janssen I, van de Kaa CH, Oosterhuis W, et al. (2005) Identification of recurrent chromosomal aberrations in germ cell tumors of neonates and infants using genomewide array-based comparative genomic hybridization. Genes Chromosomes Cancer 43: 367–76.
25. jing Sun Y, zhu Yu S, yun Sun C, Wang Q, mei Jin S, et al. (2010) [detection of chromosomal dna imbalance in medulloblastoma by comparative genomic hybridization]. Zhonghua Bing Li Xue Za Zhi 39: 606–10.
26. Cheung KJJ, Delaney A, Ben-Neriah S, Schein J, Lee T, et al. (2010) High resolution analysis of follicular lymphoma genomes reveals somatic recurrent sites of copy-neutral loss of heterozygosity and copy number alterations that target single genes. Genes Chromosomes Cancer 49: 669–81.
27. Schwaenen C, Viardot A, Berger H, Barth TFE, Bentink S, et al. (2009) Microarray-based genomic profiling reveals novel genomic aberrations in follicular lymphoma which associate with patient survival and gene expression status. Genes Chromosomes Cancer 48: 39–54.
28. Cai H, Kumar N, Baudis M (2012) arraymap: A reference resource for genomic copy number imbalances in human malignancies. PLoS One 7: 36944.
29. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466: 869–873.

DNA copy number alterations in central primitive neuroectodermal tumors and tumors of the pineal region: an international individual patient data meta-analysis

J Neurooncol (2012) 109:415–423
DOI 10.1007/s11060-012-0911-7

CLINICAL STUDY

DNA copy number alterations in central primitive neuroectodermal tumors and tumors of the pineal region: an international individual patient data meta-analysis

André O. von Bueren · Joachim Gerss · Christian Hagel · Haoyang Cai · Marc Remke · Martin Hasselblatt · Burt G. Feuerstein · Sarah Pernet · Olivier Delattre · Andrey Korshunov · Stefan Rutkowski · Stefan M. Pfister · Michael Baudis

Received: 3 April 2012 / Accepted: 4 June 2012 / Published online: 7 July 2012

Abstract Little is known about frequency, association with clinical characteristics, and prognostic impact of DNA copy number alterations (CNA) on survival in central primitive neuroectodermal tumors (CNS-PNET) and tumors of the pineal region. Searches of MEDLINE, Pubmed, and EMBASE—after the original description of comparative genomic hybridization in 1992 and July 2010—identified 15 case series of patients with CNS-PNET and tumors of the pineal region whose tumors were investigated for genome-wide CNA. One additional case study was identified from contact with experts. Individual patient data were extracted from publications or obtained from investigators, and CNAs were converted to a digitized format suitable for data mining and subgroup identification. Summary profiles for genomic imbalances were generated from case-specific data. Overall survival (OS) was estimated using the Kaplan–Meier method, and by univariable and multivariable Cox regression models. In their overall CNA profiles, low grade tumors of the pineal region clearly diverged from CNS-PNET and pineoblastoma. At a median follow-up of 89 months, 7-year OS rates of CNS-PNET, pineoblastoma, and low grade tumors of the pineal region were 22.9 ± 6, 0 ± 0, and 87.5 ± 12 %, respectively. Multivariable analysis revealed that histology (CNS-PNET), age ($\leq$2.5 years), and possibly recurrent CNAs were associated with unfavorable OS. DNA copy number profiling suggests a close relationship between CNS-PNET and pineoblastoma. Low grade tumors of the pineal region differed from CNS-PNET and pineoblastoma. Due to their

Electronic supplementary material The online version of this article (doi:10.1007/s11060-012-0911-7) contains supplementary material, which is available to authorized users.

A. O. von Bueren (✉) · S. Rutkowski
Department of Pediatric Hematology and Oncology, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, 20246 Hamburg, Germany
e-mail: a.von-bueren@uke.de

A. O. von Bueren · H. Cai · M. Baudis
Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

J. Gerss
Institute of Biostatistics and Clinical Research, University of Muenster, Muenster, Germany

C. Hagel
Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

M. Remke · S. M. Pfister
Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

M. Remke · S. M. Pfister
Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, Heidelberg, Germany

M. Hasselblatt
Institute of Neuropathology, University of Muenster, Muenster, Germany

B. G. Feuerstein
Department of Neurology, Barrow Neurological Institute—St. Joseph's Hospital and Medical Center, University of Arizona College of Medicine, Phoenix, AZ, USA

S. Pernet · O. Delattre
Institut Curie, Unité de génétique somatique, Paris, France

A. Korshunov
Department of Neuropathology, University of Heidelberg, Heidelberg, Germany

high biological and clinical variability, a coordinated prospective validation in future studies is necessary to establish robust risk factors.

Keywords Chromosomal imbalances · Prognostic markers · Comparative genomic hybridization · Brain tumor

Introduction

Central nervous system primitive neuroectodermal tumors (CNS-PNET) are a heterogeneous group of WHO grade IV lesions (Supplementary Table 1). They comprise 3–7 % of brain tumors in children and young adults [1, 2] and are associated with a dismal prognosis [3, 4]. Histologically, these highly proliferative lesions are currently divided into CNS-PNET or supratentorial PNET, respectively (synonym PNET not otherwise specified, PNET NOS), CNS neuroblastoma, CNS ganglioneuroblastoma, medulloepithelioma, and ependymoblastoma [5]. CNS-PNET and medulloblastoma share a similar histology and are often solely distinguishable by their supratentorial versus infratentorial location. Further, pineoblastoma, a WHO grade IV tumor of the pineal gland [5], is filed in some studies as CNS-PNET although pineoblastoma forms a group of neoplasms of the pineal region together with pineocytoma, pineal parenchymal tumor of intermediate differentiation, and papillary tumor of the pineal region [5]. The classification of malignancies within the group of embryonal tumors has changed considerably in the last four editions of the WHO classification of tumors of the CNS (Supplementary Table 1). Tumor classification systems are increasingly complemented by molecular genetic profiling data, especially in hematologic neoplasias [6]. However, for the various subtypes of CNS-PNET, such data are still scarce and large series are missing. Profiling of regional copy number abnormalities (CNA) by genomic hybridization techniques is a robust methodology for whole genome data analysis. Principal techniques include the different variants of chromosomal and array-based comparative genomic hybridization (cCGH/aCGH; [7–10]) and single-color oligonucleotide array technologies [e.g., genomic single nucleotide polymorphism (SNP) arrays].

In contrast to data from gene expression measurements, CGH data is easily adaptable across multiple datasets to perform a meta-analysis. Methods to assess genomic CNAs are standardized and reproducible as demonstrated in previous reports (e.g., [11, 12]). Some earlier reviews have reported on specific types of aberrations or were focused on the descriptive analysis of certain classes of malignancies [13, 14].

Due to the low incidence of CNS-PNET and pineoblastoma, only a few CGH studies have been reported in these tumors [2, 15–17]. So far, results have suggested that CNS-PNET are genetically heterogeneous with frequent and diverse CNAs and that CNA patterns are distinct from those observed in medulloblastoma [2, 15–17].

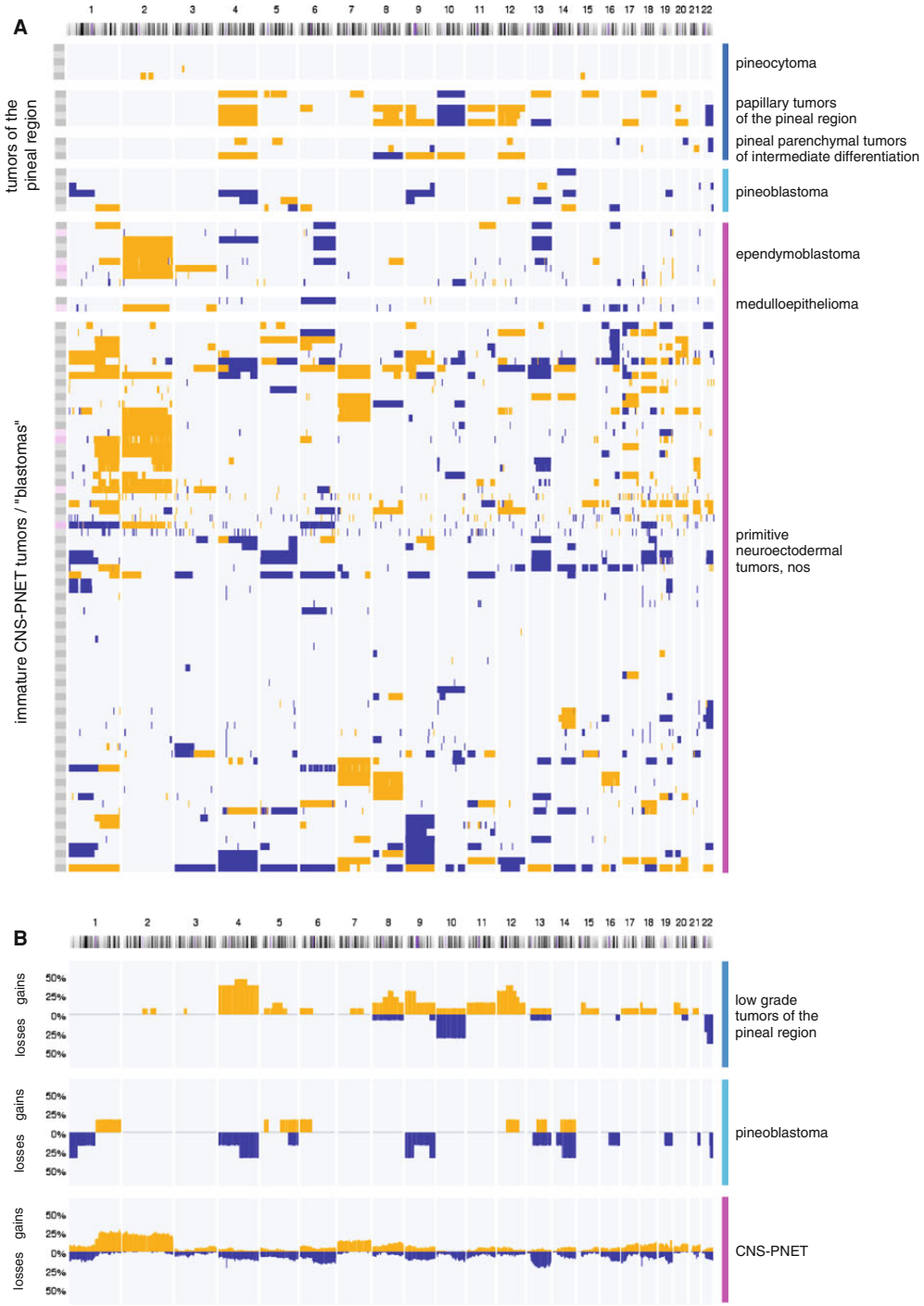
Fig. 1 Delineation of 3 distinct clinicogenetic subgroups. **a** Regional copy number imbalances for individual cases were plotted separately by overall diagnostic assignment [*yellow* gain, *blue* loss, *blue* tumors of the pineal region except pineoblastoma, *light blue* pineoblastoma, *pink* central primitive neuroectodermal tumors (CNS-PNET)]. Individual profiles were arranged by hierarchical clustering inside their groups. **b** Histograms of genomic gain and loss frequencies (*color legend* corresponding to (**a**))

For the present study, we performed an individual patient data (IPD) meta-analysis—a specific method of systematic review [18] offering advantages for meta-analysis [19, 20]—of genomic imbalances in CNS-PNET and tumors of the pineal region. The collected data are made available through the "Progenetix" molecular-cytogenetic database (www.progenetix.org: [14, 21, 22]).

Methods

Search strategy, and selection criteria

We did a modification of the Cochrane Highly Sensitive Search Strategy for prognostic studies [20] combined with predefined search terms in MEDLINE, Pubmed, and EMBASE without language restriction [23, 24]. The process of the study retrieval, in- and exclusion of studies/patients is displayed in the flow chart (Supplementary Fig. 1) according to the PRISMA (preferred reporting items for systematic reviews and meta-analyses) statement. The search was limited to articles published after the original description of CGH [7] until July 2010. Key words were: "medullo(-)blastoma(s)", "primitive neuroectodermal tumo(u)r(s)", "neuroectodermal tumo(u)r(s) primitive" "pnet (s)", "medullo(-)epithelioma(s)", "ependymoblastoma(s)", "ganglioneuroblastoma(s)", "pinealoma", "pineocytoma(s)", "pineoblastoma(s)", "pineal tumo(u)r(s)", "pineal parenchymal tumo(u)r(s)", "mixed transitional pineal tumo(u)r(s)", "mixed transitional pineal tumo(u)r(s)", "atypical teratoid rhabdoid tumo(u)r(s)", "rhabdoid tumo(u)r(s)", "AT(/)RT" and "rhabdoid", "supratentorial neoplasm(s)" or "neuroblastoma(s)" and "central nervous system neoplasm(s)"; and "cgh" or "comparative genomic hybridization" or "snp" or "SNP" or "genomic array(s)" or "copy number" or "dna microarray(s)" or "amplification". Additionally to the search queries, we followed references from the selected articles and assessed each abstract. Minimal requirements for inclusion of a patient to the study were the availability of case-specific genomic copy number data with whole genome coverage, the unambiguous diagnostic classification of CNS-PNET/tumor of the pineal region, and matching available or inferred locus information.


A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
tumors of the pineal region
immature CNS-PNET tumors / "blastomas"
pineocytoma
papillary tumors of the pineal region
pineal parenchymal tumors of intermediate differentiation
pineoblastoma
ependymoblastoma
medulloepithelioma
primitive neuroectodermal tumors, nos
B
gains
losses
50%
25%
0%
low grade tumors of the pineal region
pineoblastoma
CNS-PNET

Clinical and CNA data collection, data extraction, quality assessment, conversion of CNA data, and data synthesis

For CGH results specified in cytogenetic annotation formats, data were standardized to ISCN 1995 (International System for Human Cytogenetic Nomenclature, 1995) "revish" format based on an 862-bands karyotype and checked for semantically correct annotation using dedicated software. For genomic array data without annotated gain/loss information, clone specific data files were segmented using Progenetix website tools. Normalized data were converted to Golden Path mapped copy number status information by software implemented in the Perl scripting language [14].

In a first step, clinical and genomic data were extracted from publications by two reviewers (A.O.V.B. and M.B.). Subsequently, the original data, in particular in case of incomplete data (genomic and clinical data), for each participant were obtained and updated directly from the researcher responsible for each included study [25]. To prevent duplicate inclusions, authors were asked to indicate whether a patient had been analyzed within different studies. In addition, copy number profiles were clustered for similarity and reviewed for the occurrence of profile pairs, in order to avoid duplicate cases due to republished data. Data of three unpublished CNS-PNET patients were provided by two authors (S.P. and O.D.). Generally, two approaches to perform IPD meta-analyses are used. First, IPD meta-analyses can be performed directly, as if all data belong to a single trial/study, termed the "one-stage" approach [26]. Second, a "two-stage" approach can also be used. Each trial/study is analyzed separately using its raw data before the summary results from each trial/study are pooled and analyzed using conventional meta-analyses techniques [26]. Due to the small patient numbers of each individual case series, the "one-stage" approach was used here.

Exploratory data mining and statistical analysis

For the evaluation of regional copy number changes, non-overlapping genomic segments were generated based on the complete CNA data from all cases. For each of these intervals, case-specific involvement was evaluated and gain/loss frequencies determined. For visualization and ordering of case-specific CNA data, data matrices were produced containing imbalance status (gain, or loss) mapped to a variety of genomic intervals (from chromosomal arm level down to 1 Mb). Cases were ordered by hierarchical clustering of gain/loss matrices (unsupervised, complete linkage), and the derived case order was used for re-plotting of the original CNA annotations. CNA complexity, a relatively resolution-independent surrogate marker of genomic instability, was determined for each case by evaluating the occurrence of gain and loss events per chromosome arm, with a maximum score of 2 per arm (i.e. occurrence of one or more of each gain and loss; modified from [27]).

To evaluate imbalance distribution in relation to diagnostic assignment, for each of the entities in our dataset, gain/loss frequencies were calculated mapped to genomic intervals on a 5-Mb level. Copy number profiles were compared by generating a heatmap of gain/loss distributions.

Cases with clinical follow-up were evaluated with respect to correlation of clinical factors and regional CNA status to OS. OS was defined as date of diagnosis to death of any cause or to the date of last visit. Cut-off values of age and CNA complexity were determined by recursive partitioning [28]. Univariable and multivariable survival analyses were performed. OS was estimated by the Kaplan–Meier method, and the log-rank test was used for comparisons of survival in different groups [29]. Univariable analyses to investigate the effect of age (continuous), and CNA complexity (continuous) on OS was done with univariable Cox regression analysis. Multivariable analyses were performed using Cox's proportional hazards model. All statistical analyses are intended to be rather exploratory than confirmatory. p values are considered statistically significant when $p < 0.05$. No adjustment for multiple testing was carried out. Statistical analyses were performed using SAS (v.9.2 for Windows; SAS Institute, Cary, NC, USA), and PASW Statistics 18 for Windows (SPSS, Chicago, IL, USA).

Results

Supplementary Figure 1 illustrates the process of evaluating articles for inclusion in the IPD meta-analysis. We identified 1,220 papers by the search terms. The number of papers was reduced to 840 after removing of duplicates (by titles and abstracts). Title and abstract review resulted in the exclusion of 710 papers. Three case-specific data (one case series) were provided by two authors. We reviewed 131 papers in full, from which 15 studies, and 1 unpublished case series ($n = 3$), met inclusion criteria for this study (Supplementary Fig. 1).

Study characteristics and quality assessment

The 16 studies included here comprised 107 patients in total, after exclusion of 4 cases with ambiguous CNA profiles. From 61 patients, information about OS was available (clinical characteristics are shown in Table 1). Of those, 38 patients were profiled using aCGH and 23 patients using cCGH. The median follow-up time for

Table 1 Demographics and disease characteristics of 61 patients with central primitive neuroectodermal tumors (*CNS-PNET*) and tumors of the pineal region

Characteristics	Number of patients (complete follow-up; $n = 61$)
Sex	
Male	13 (21 %)
Female	17 (28 %)
N/A	31 (51 %)
Age	
Median age at diagnosis (range; years)	4.2 (0.6–66)
Histology	
CNS-PNET	46 (75 %)
Tumors of the pineal region	15 (25 %)
Tumor samples source	
Primary tumors	59 (97 %)
Relapses	2 (3 %)
Metastatic stage	
Metastases	8 (13 %)
No metastases	21 (35 %)
N/A	32 (52 %)

N/A information not available

survivors was 75 months, and the median follow-up time across all patients was 89 months. Fifteen children were aged $\leq$2.5 years and 46 patients were aged >2.5 years. The cohort compromised all tumor entities classified as CNS-PNET in the current WHO classification when taking into account the update of earlier WHO classification in which some of these tumors were partly classified as different subgroups of embryonal tumors [5, 30] ($n = 46$), and tumors of the pineal region ($n = 15$) which included pineocytoma ($n = 4$), pineal parenchymal tumor of intermediate differentiation ($n = 3$), papillary tumor of the pineal region ($n = 5$), and pineoblastoma ($n = 3$). Mean CNA complexity was 9.4 (range, 0.00–30.00). For the purpose of statistical analysis, CNS-PNET were considered as one group and tumors of the pineal region were considered as another group.

Overall genomic imbalance patterns in central nervous system primitive neuroectodermal tumors and tumors of the pineal region

In order to evaluate the overall patterns of genomic imbalances in bona fide CNS-PNET and tumors of the pineal region, we visualized the case-specific CNAs of all tumors clustered for their overall imbalance similarities (Fig. 1a). In CNS-PNET ($n = 88$), frequent gains of chromosomes 1q4 [$n = 31$ (35 %)], 2p2 [$n = 27$ (31 %)], and 7q3 [$n = 16$ (18 %)] as well as losses involving chromosome 13q2 [$n = 21$ (24 %)], and 6q [$n = 18$ (20 %)] could be observed among other less frequent changes (Fig. 1b). In contrast, low grade tumors of the pineal region were characterized by gains of chromosomes 4q2 [$n = 6$ (46 %)], and 12 [$n = 5$ [38 %)] as well as losses of chromosomes 10 [$n = 4$ (31 %)), and 22 [$n = 5$ (38 %)]. Interestingly, pineoblastoma ($n = 6$) displayed a pattern of genomic imbalances unrelated to the changes observed in the group of low grade tumors of the pineal region. Supplementary Figs. 2–4 illustrate gains and losses of the different disease entities.

We observed frequent gains involving chromosome 2 and losses involving chromosome 6 in ependymoblastoma as well as in medulloepithelioma (Supplementary Fig. 3b, c). Losses of chromosome 6 and 13 were typical for ependymoblastoma.

Embryonal tumor with abundant neuropil and true rosettes (ETANTR) was first described by Eberhart et al. [31], but is so far not listed as a distinct tumor entity in the 2007 WHO classification [5] and represents a CNS-PNET with "ependymoblastic" rosettes [32]. Recently, Korshunov et al. [33] demonstrated in a series of 21 ependymoblastoma and 20 ETANTR that 95 % of ETANTRs and 90 % of ependymoblastoma have the unique focal amplification at 19q13.42.

Therefore, the term embryonal tumor with multilayered rosette (ETMR) has been suggested for ependymoblastoma and ETANTR, a new entity with multilayered rosettes for which amplification at 19q13.42 represents a rather sensitive and specific marker [32].

In our cohort, we identified 9 tumors with such an amplification. As described previously by Li et al. [2], cases with such an amplification predominantly (8/9) also displayed gains of the whole or the major part of chromosome 2. For some additional cases with gain of chromosome 2 identified by cCGH, no high-resolution data were available. Therefore, we may not rule out an additional amplification at 19q13.42 in these cases.

Univariable and multivariable survival analysis of clinical factors and CNA complexity

To assess which parameters contribute to prognosis, we evaluated each clinical variable by univariable Kaplan–Meier analysis. Tested variables were: gender, age, histology (CNS-PNET vs. tumors of the pineal region), metastatic stage (no metastases vs. metastases), extent of postoperative residual disease (complete/gross total resection vs. residual disease $\geq$1.5 cm^2), radiotherapy (no radiotherapy/local radiotherapy vs. cranio-spinal radiotherapy), chemotherapy (no chemotherapy vs. chemotherapy), CNA complexity (<11 vs $\geq$11 as defined by recursive partitioning), tumor

sample source (primary tumor vs. relapse), and technique (aCGH vs. cCGH). Supplementary Table 2 illustrates the factors (histology, CNA complexity, and age) showing differences as assessed by univariable analysis. Patients with tumors of the pineal region had a more favorable OS when compared to patients with CNS-PNET (7-year OS: 64.7 ± 15 vs. 22.9 ± 6 %, $p = 0.007$). Of note, all three patients with a pineoblastoma and available follow-up were dead 33 months after diagnosis, whereas all other patients with low grade tumors of the pineal region had excellent outcome (7-year OS: 87.5 ± 12 %). Patients aged $\leq$2.5 years had unfavorable OS when compared to patients aged >2.5 years (7-year OS: 0 ± 0 vs. 41.3 ± 8 %, $p = 0.001$). OS rates were similar in CNS-PNET patients with and without the amplification at 19q13.42. Univariable cox regression analysis confirmed that increasing age (continuous variable) is denoting a more favorable OS [hazard ratio, 0.967 (per year); 95 % confidence interval, 0.939–0.996; $p = 0.0282$] and increasing CNA complexity (continuous variable) a less favorable OS [hazard ratio, 1.063 (per unit); 95 % confidence interval, 1.012–1.117; $p = 0.0153$]. Multivariable analysis of clinical factors and CNA complexity revealed that histology (tumors of the pineal region), age (older than 2.5 years) and CNA complexity <11 are favorable prognostic factors (Table 2).

Multivariable survival analysis of chromosomal aberrations, CNA complexity, and clinical factors

To identify which of the chromosomal aberrations might have an impact on OS, multivariable survival analyses were applied to all 61 patients incorporating the significant clinical factors (histology and age), CNA complexity, as well as 75 different chromosomal gains and 75 different chromosomal losses in a stepwise approach, respectively. These analyses finally revealed that young age ($\leq$2.5 years), histology (CNS-PNET), and recurrent gains of 3p1 ($n = 3$; 5 %), 13q1 ($n = 5$; 8.2 %), and 15q2 ($n = 8$; 13.1 %) are associated with an increased risk for unfavorable OS (Table 3).

Discussion

Over recent years, whole genome/transcriptome molecular analysis has led to the identification of divergent biological characteristics in what were considered single cancer types. In the field of pediatric neuro-oncology, medulloblastoma are now considered as a group of biologically differing entities consisting of at least 4 molecular subgroups, loosely connected through their topography (cerebellum) and partially overlapping histological appearance [34–42].

Molecular studies in rare tumor entities are severely limited due to the low number of cases included in single series, as well as conceptual and technical heterogeneity of the studies. To our knowledge, our study is the first IPD meta-analysis assessing the genomic and clinical features in CNS-PNET and tumors of the pineal region and their impact on OS. In this study, we show that CNS-PNET and pineoblastoma are divergent in their CNA profiles when compared with low grade tumors of the pineal region. For the cases analyzed here, recurring CNA observed only in low grade tumors of the pineal region were, e.g., gains on

Table 2 Multivariable analyses of clinical prognostic factors ($n = 61$) for overall survival (OS)

Parameter	Comparison	Sample size	HR OS	95 % Confidence interval	*P* value
Histology	Non CNS-PNET	15	0.312	0.109–0.891	0.0296
	CNS-PNET	46			
Age group (years)	>2.5	46	0.386	0.197–0.757	0.0056
	≤2.5	15			
CNA complexity	≥11 CNA	23	1.790	0.943–3.400	0.0752
	<11 CNA	38			

CNS-PNET Central primitive neuroectodermal tumor, *Non CNS-PNET* tumors of the pineal region, *CNA* copy number aberrations, *HR OS* Hazard ratio overall survival

Table 3 Multivariable analyses of clinical factors and recurrent chromosomal aberrations (forward stepwise selection; $n = 61$) for overall survival

Parameter	Sample size	Hazard ratio overall survival	95 % Confidence interval	*P* value
Age (≥2.5 years)	46	0.295	0.141–0.619	0.0012
Histology (tumor of the pineal region)	15	0.120	0.029–0.498	0.0035
seg3p1_gain	3	8.759	1.778–43.159	0.0077
seg13q1_gain	5	4.128	1.192–14.303	0.0253
seg15q2_gain	8	4.338	1.614–11.665	0.0036

4q2, 9p, 12p, and 8q2 as well as deletions of chromosome 10. In contrast, recurring CNA only found in pineoblastoma were deletions on 4q, chromosome 9, and 1p3. Based on our results, CGH analysis might be of help—in addition to neuroradiological and histopathological evaluation—to differentiate between CNS-PNET, pineoblastoma, and lower WHO grade tumors of the pineal region. While detection of the listed aberrations may be indicative for assignment to one of the diagnostic groups, development of a CNA-based classifier will ideally require larger numbers of genome profiles.

We found evidence that younger age at time of diagnosis is a negative prognostic factor for OS, confirming several previous studies reporting on poor outcome of young children with CNS-PNET/pineoblastoma [3, 43]. Timmermann et al. [3] reported on OS and progression-free survival rates after 3 years of 17.2 and 14.9 %, respectively. Administration of radiotherapy was the only significant prognostic marker (15 out of 29 patients were not irradiated) in this study [3] suggesting that omitting the radiotherapy in young children—with the goal to reduce neurologic sequelae—might at least explain partly the extremely poor outcome of young children with CNS-PNET/pineoblastoma.

In our cohorts, CNS-PNET and pineoblastoma shared an unfavorable prognosis. Small numbers of pineoblastoma (3 out of 61 patients) may limit the comparison of those two tumor entities. Based on the literature, there is some evidence that patients with pineoblastoma may do better than patients with CNS-PNET [44, 45]. Patients with low grade tumors of the pineal region had a favorable outcome (7-year OS: 87.5 ± 12 %) confirming that those tumor entities need a less aggressive treatment than CNS-PNET/pineoblastoma.

CNS-PNET and tumors of the pineal region share a complex karyotype with frequent CNAs [46]. In our series of 107 patients, low grade tumors of the pineal region showed relative frequently absence of CNAs (4/13), less frequently in pineoblastoma (1/6), and CNS-PNET (2/88).

Recently, a new entity of CNS-PNET termed ETMR has been suggested for a subgroup of CNS-PNET (ependymoblastoma and ETANTR) for which amplification at 19q13.42 represents a rather sensitive and specific marker [32]. Korshunov et al. [33] identified in the great majority of ependymoblastoma and ETANTR the focal amplification at 19q13.42 whereas such an amplification was not observed in a large series of other pediatric brain tumors [32]. As we report about cCGH and aCGH data, the frequency of tumors with amplification at 19q13.42 (Supplementary Fig. 6) should be interpreted with caution as detection of the amplification at 19q13.42 might be missed when tumors are profiled by conventional cCGH, which has a spatial resolution limited of several megabases. Patients with 19q13.42 amplified tumors had a relatively poor OS (6/7 patients with available follow-up died of disease). Of note, the analysis of the prognostic impact of the amplification at 19q13.42 is limited in our cohort, because—as mentioned above—this amplification might be missed in tumors analyzed with cCGH.

Our results provide evidence that high CNA complexity is an unfavorable prognostic marker in our cohort. Because of high frequencies of genomic imbalances as well as heterogeneous patterns and frequencies of CNAs, CNA complexity appears to be a good measure for overall genomic instability which may reflect aggressiveness of a certain tumor. In light of this, specific recurrent genomic imbalances which have been identified as CNAs with potential impact on OS in our analyses [e.g., in the 61 patients: gain of seg3p1 ($n = 3$), seg13q1 ($n = 5$), seg15q2 ($n = 8$)], need to be validated—ideally in large future studies—for their prognostic value.

After the search cut-off date imposed by the IPD meta-analysis criteria, another study was published recently focusing on CNS-PNET/pineoblastoma only in pediatric patients [17]. By evaluating the genomic array data which are available from NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; accession number GSE12370), we were able to generate CNA profiles for 38 patients (8 of whom had pineoblastoma, and 30 had a CNS-PNET; CGH data from 35 CNS-PNET cases were listed, 5 recurrent tumors were paired with a primary sample from the same patient) and 1 CNS-PNET cell line. Here, as in our IPD meta-analysis, pineoblastoma exhibited CNA profiles roughly comparable to subsets of cases identified as CNS-PNET as shown in the Supplementary Fig. 5 a, b.

The approach of an IPD meta-analysis—a specific method of systematic review based on a systematic search—is in our opinion both necessary and efficient to increase the patient number in rare tumor diseases. By using IPD, we may overcome many of the limitations of systematic reviews (e.g., poor quality of data can be improved by updating the information). We used common inclusion and exclusion criteria for each individual case. In addition, we have performed a quality assessment of genomic data by reassessment of each individual case by two researchers (M.B. and H.C.). Methods to assess genomic CNAs are standardized and reproducible as demonstrated in previous reports (e.g., [11, 12]). Moreover, by including unpublished data [25], we aimed to reduce the risk for publication bias [20]. Of course, the inclusion of a larger number of unpublished cases would have been desirable, and a "pooling" of such data has an exceptional value for rare diseases. Of note, IPD meta-analyses usually take longer than conventional systematic review, and obtaining IPD is time-consuming [20]. Therefore, it is not possible to include all very recent studies, and many IPD

meta-analyses are conducted on a cyclical basis with data collection, quality assessment, analyses, and dissemination of results taking place every few years [18], because by the time of the final analysis of the pooled data new cases are already available. We acknowledge some limitations of our study which is based on original data produced over a time period of several years. As shown in Supplementary Table 1, the WHO classification of tumors of the CNS has changed during this period. Moreover, in recent years, the staging has improved, as have surgical procedures and non-surgical treatment options of patients with CNS-PNET and tumors of the pineal region. Regarding genomic analysis methods, high-resolution profiling by genomic copy number arrays or whole genome sequencing could provide a higher sensitivity for the detection of hitherto undetected CNA. However, the main limitations in identifying robust CNA markers with prognostic value are in the limited number of samples and associated clinical datasets available for such analyses.

In summary, CNS-PNET and low grade tumors of the pineal region are characterized by differences in CNA profiles. In this respect, pineoblastoma fit readily into the genomically heterogeneous group of CNS-PNET with a complex karyotype. Although not necessarily displayed by each individual case, typical CNA profiles underline the differing biological background of these entities. Our results provide evidence that young age, high CNA complexity, and potentially also several specific CNAs may have an impact on OS.

Acknowledgments The authors' are indebted to the authors of articles, who provided the data to this study that otherwise would not have been accessible. In particular, the authors would like to thank the following researcher/clinicians for their help: Milo Puhan, Carolyn Russo, Wolfram Scheurlen, Barbara Schütz, Christine Haberler, Martin McCabe, and Hans-Hermann Dubben. The author would like to thank Klaus-Dieter Papke for assisting the literature search. The authors acknowledge the following sources of funding: German Children's Cancer Foundation/Deutsche Kinderkrebsstiftung (to A.O.V.B., S.R.). Haoyang Cai is supported through a grant from the China Scholarship Council.

Conflict of interest None.

References

1. Gaffney CC, Sloane JP, Bradley NJ, Bloom HJ (1985) Primitive neuroectodermal tumours of the cerebrum. Pathology and treatment. J Neurooncol 3:23–33
2. Li M, Lee KF, Lu Y, Clarke I, Shih D, Eberhart C, Collins VP, Van Meter T, Picard D, Zhou L, Boutros PC, Modena P, Liang ML, Scherer SW, Bouffet E, Rutka JT, Pomeroy SL, Lau CC, Taylor MD, Gajjar A, Dirks PB, Hawkins CE, Huang A (2009) Frequent amplification of a chr19q13.41 microRNA polycistron in aggressive primitive neuroectodermal brain tumors. Cancer Cell 16:533–546
3. Timmermann B, Kortmann RD, Kuhl J, Rutkowski S, Meisner C, Pietsch T, Deinlein F, Urban C, Warmuth-Metz M, Bamberg M (2006) Role of radiotherapy in supratentorial primitive neuroectodermal tumor in young children: results of the German HIT-SKK87 and HIT-SKK92 trials. J Clin Oncol 24:1554–1560
4. Fangusaro J, Massimino M, Rutkowski S, Gururangan S (2010) Non-cerebellar primitive neuroectodermal tumors (PNET): summary of the Milan consensus and state of the art workshop on marrow ablative chemotherapy with hematopoietic cell rescue for malignant brain tumors of childhood and adolescents. Pediatr Blood Cancer 54:638–640
5. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK (2007) WHO classification of tumours of the central nervous system. IARC, Lyon
6. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, Harris NL, Le Beau MM, Hellstrom-Lindberg E, Tefferi A, Bloomfield CD (2009) The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood 114:937–951
7. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258:818–821
8. Joos S, Bergerheim US, Pan Y, Matsuyama H, Bentz M, du Manoir S, Lichter P (1995) Mapping of chromosomal gains and losses in prostate cancer by comparative genomic hybridization. Genes Chromosom Cancer 14:267–276
9. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosom Cancer 20:399–407
10. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41–46
11. Bown N, Cotterill S, Lastowska M, O'Neill S, Pearson AD, Plantaz D, Meddeb M, Danglot G, Brinkschmidt C, Christiansen H, Laureys G, Speleman F, Nicholson J, Bernheim A, Betts DR, Vandesompele J, Van Roy N (1999) Gain of chromosome arm 17q and adverse outcome in patients with neuroblastoma. N Engl J Med 340:1954–1961
12. Zenz T, Mertens D, Dohner H, Stilgenbauer S (2008) Molecular diagnostics in chronic lymphocytic leukemia—pathogenetic and clinical implications. Leuk Lymphoma 49:864–873
13. Moinzadeh P, Breuhahn K, Stutzer H, Schirmacher P (2005) Chromosome alterations in human hepatocellular carcinomas correlate with aetiology and histological grade–results of an explorative CGH meta-analysis. Br J Cancer 92:935–941
14. Baudis M (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. BMC Cancer 7:226
15. Russo C, Pellarin M, Tingby O, Bollen AW, Lamborn KR, Mohapatra G, Collins VP, Feuerstein BG (1999) Comparative genomic hybridization in patients with supratentorial and infratentorial primitive neuroectodermal tumors. Cancer 86:331–339
16. Pfister S, Remke M, Toedt G, Werft W, Benner A, Mendrzyk F, Wittmann A, Devens F, von Hoff K, Rutkowski S, Kulozik A, Radlwimmer B, Scheurlen W, Lichter P, Korshunov A (2007) Supratentorial primitive neuroectodermal tumors of the central nervous system frequently harbor deletions of the CDKN2A locus and other genomic aberrations distinct from medulloblastomas. Genes Chromosom Cancer 46:839–851
17. Miller S, Rogers HA, Lyon P, Rand V, Adamowicz-Brice M, Clifford SC, Hayden JT, Dyer S, Pfister S, Korshunov A, Brundler MA, Lowe J, Coyle B, Grundy RG (2011) Genome-wide

molecular characterization of central nervous system primitive neuroectodermal tumor and pineoblastoma. Neuro Oncol 13:866–879
18. Clarke M, Godwin J (1998) Systematic reviews using individual patient data: a map for the minefields? Ann Oncol 9:827–833
19. Riley RD, Sauerbrei W, Altman DG (2009) Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. Br J Cancer 100:1219–1229
20. Altman DG (2001) Systematic reviews of evaluations of prognostic variables. BMJ 323:224–228
21. Baudis M, Cleary ML (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 17:1228–1229
22. Baudis M (2006) Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. Biotechniques 40:269–270, 272
23. Wilne S, Collier J, Kennedy C, Koller K, Grundy R, Walker D (2007) Presentation of childhood CNS tumours: a systematic review and meta-analysis. Lancet Oncol 8:685–695
24. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6:e1000097
25. Stewart LA, Tierney JF (2002) To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof 25:76–97
26. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG (2005) Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clin Trials 2:209–217
27. Boerma EG, Siebert R, Kluin PM, Baudis M (2009) Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge. Leukemia 23:225–234
28. LeBlanc M, Crowley J (1992) Relative risk trees for censored survival data. Biometrics 48:411–425
29. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
30. Kleihues P, Cavenee WK (2000) World Health Organization classification of tumours. Pathology and genetics of tumours of the nervous system. IARC, Lyon
31. Eberhart CG, Brat DJ, Cohen KJ, Burger PC (2000) Pediatric neuroblastic brain tumors containing abundant neuropil and true rosettes. Pediatr Dev Pathol 3:346–352
32. Paulus W, Kleihues P (2010) Genetic profiling of CNS tumors extends histological classification. Acta Neuropathol 120:269–270
33. Korshunov A, Remke M, Gessi M, Ryzhova M, Hielscher T, Witt H, Tobias V, Buccoliero AM, Sardi I, Gardiman MP, Bonnin J, Scheithauer B, Kulozik AE, Witt O, Mork S, von Deimling A, Wiestler OD, Giangaspero F, Rosenblum M, Pietsch T, Lichter P, Pfister SM (2010) Focal genomic amplification at 19q13.42 comprises a powerful diagnostic marker for embryonal tumors with ependymoblastic rosettes. Acta Neuropathol 120:253–260
34. Schwalbe EC, Lindsey JC, Straughton D, Hogg TL, Cole M, Megahed H, Ryan SL, Lusher ME, Taylor MD, Gilbertson RJ, Ellison DW, Bailey S, Clifford SC (2011) Rapid diagnosis of medulloblastoma molecular subgroups. Clin Cancer Res 17:1883–1894
35. Thompson MC, Fuller C, Hogg TL, Dalton J, Finkelstein D, Lau CC, Chintagumpala M, Adesina A, Ashley DM, Kellie SJ, Taylor MD, Curran T, Gajjar A, Gilbertson RJ (2006) Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. J Clin Oncol 24:1924–1931
36. Kool M, Koster J, Bunt J, Hasselt NE, Lakeman A, van Sluis P, Troost D, Meeteren NS, Caron HN, Cloos J, Mrsic A, Ylstra B, Grajkowska W, Hartmann W, Pietsch T, Ellison D, Clifford SC, Versteeg R (2008) Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PLoS One 3:e3088
37. Cho YJ, Tsherniak A, Tamayo P, Santagata S, Ligon A, Greulich H, Berhoukim R, Amani V, Goumnerova L, Eberhart CG, Lau CC, Olson JM, Gilbertson RJ, Gajjar A, Delattre O, Kool M, Ligon K, Meyerson M, Mesirov JP, Pomeroy SL (2011) Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J Clin Oncol 29:1424–1430
38. Northcott PA, Korshunov A, Witt H, Hielscher T, Eberhart CG, Mack S, Bouffet E, Clifford SC, Hawkins CE, French P, Rutka JT, Pfister S, Taylor MD (2011) Medulloblastoma comprises four distinct molecular variants. J Clin Oncol 29:1408–1414
39. Remke M, Hielscher T, Northcott PA, Witt H, Ryzhova M, Wittmann A, Benner A, von Deimling A, Scheurlen W, Perry A, Croul S, Kulozik AE, Lichter P, Taylor MD, Pfister SM, Korshunov A (2011) Adult medulloblastoma comprises three major molecular variants. J Clin Oncol 29:2717–2723
40. Remke M, Hielscher T, Korshunov A, Northcott PA, Bender S, Kool M, Westermann F, Benner A, Cin H, Ryzhova M, Sturm D, Witt H, Haag D, Toedt G, Wittmann A, Schottler A, von Bueren AO, von Deimling A, Rutkowski S, Scheurlen W, Kulozik AE, Taylor MD, Lichter P, Pfister SM (2011) FSTL5 is a marker of poor prognosis in non-WNT/non-SHH medulloblastoma. J Clin Oncol 29:3852–3861
41. Taylor MD, Northcott PA, Korshunov A, Remke M, Cho YJ, Clifford SC, Eberhart CG, Parsons DW, Rutkowski S, Gajjar A, Ellison DW, Lichter P, Gilbertson RJ, Pomeroy SL, Kool M, Pfister SM (2012) Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol 123:465–472
42. Kool M, Korshunov A, Remke M, Jones DT, Schlanstein M, Northcott PA, Cho YJ, Koster J, Schouten-van Meeteren A, van Vuurden D, Clifford SC, Pietsch T, von Bueren AO, Rutkowski S, McCabe M, Collins VP, Backlund ML, Haberler C, Bourdeaut F, Delattre O, Doz F, Ellison DW, Gilbertson RJ, Pomeroy SL, Taylor MD, Lichter P, Pfister SM (2012) Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. Acta Neuropathol 123:473–484
43. Geyer JR, Sposto R, Jennings M, Boyett JM, Axtell RA, Breiger D, Broxson E, Donahue B, Finlay JL, Goldwein JW, Heier LA, Johnson D, Mazewski C, Miller DC, Packer R, Puccetti D, Radcliffe J, Tao ML, Shiminski-Maher T (2005) Multiagent chemotherapy and deferred radiotherapy in infants with malignant brain tumors: a report from the Children's Cancer Group. J Clin Oncol 23:7621–7631
44. Pizer BL, Weston CL, Robinson KJ, Ellison DW, Ironside J, Saran F, Lashford LS, Tait D, Lucraft H, Walker DA, Bailey CC, Taylor RE (2006) Analysis of patients with supratentorial primitive neuro-ectodermal tumours entered into the SIOP/UKCCSG PNET 3 study. Eur J Cancer 42:1120–1128
45. Timmermann B, Kortmann RD, Kuhl J, Meisner C, Dieckmann K, Pietsch T, Bamberg M (2002) Role of radiotherapy in the treatment of supratentorial primitive neuroectodermal tumors in childhood: results of the prospective German brain tumor trials HIT 88/89 and 91. J Clin Oncol 20:842–849
46. Li MH, Bouffet E, Hawkins CE, Squire JA, Huang A (2005) Molecular genetics of supratentorial primitive neuroectodermal tumors and pineoblastoma. Neurosurg Focus 19:E3

BIBLIOGRAPHY

[1] American Cancer Society. Global Cancer Facts & Figures 2nd Edition. Atlanta: American Cancer Society; 2011.

[2] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646-74.

[3] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009 Apr 9;458(7239):719-24.

[4] Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, et al. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546-58.

[5] von Hansemann, D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. Virchows Arch. Path. Anat. 1890; 119, 299.

[6] Boveri, T. Zur Frage der Entstehung Maligner Tumoren. Gustav Fischer 1914; 1–64.

[7] Avery, O. T., MacLeod, C. M. and McCarty, M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J. Exp. Med. 1944; 79, 137–158.

[8] Watson, J. D. and Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953; 171, 737–738.

[9] Nowell, P. and Hungerford, D. A minute chromosome in human granulocytic leukemia. Science. 1960; 132, 1497.

[10] Rowley, J. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature. 1973; 243, 290–293.

[11] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992 Oct 30;258(5083):818-21.

[12] Joos S, Scherthan H, Speicher MR, Schlegel J, et al. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993 Feb;90(6):584-9.

[13] du Manoir S, Speicher MR, Joos S, Schröck E, et al. Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Hum Genet. 1993 Feb;90(6):590-610.

[14] Wade M, Li YC, Wahl GM. MDM2, MDMX and p53 in oncogenesis and cancer therapy. Nat Rev Cancer. 2013 Feb;13(2):83-96.

[15] Wirtz D, Konstantopoulos K, Searson PC. The physics of cancer: the role of physical interactions and mechanical forces in metastasis. Nat Rev Cancer. 2011 Jun 24;11(7):512-22.

[16] Bentzen SM. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. Nat Rev Cancer. 2006 Sep;6(9):702-13.

[17] Chabner BA, Roberts TG Jr. Timeline: Chemotherapy and the war on cancer. Nat Rev Cancer. 2005 Jan;5(1):65-72.

[18] Palucka K, Banchereau J. Cancer immunotherapy via dendritic cells. Nat Rev Cancer. 2012 Mar 22;12(4):265-77.

[19] Ataman OU, Sambrook SJ, Wilks C, Lloyd A, et al. The clinical development of molecularly targeted agents in combination with radiation therapy: a pharmaceutical perspective. Int J Radiat Oncol Biol Phys. 2012 Nov 15;84(4):e447-54.

[20] Wang LC, Wang L, Kwauk S, Woo JA, et al. Analysis on the clinical features of 22 basaloid squamous cell carcinoma of the lung. J Cardiothorac Surg. 2011 Jan 26;6:10.

[21] Subramanian J, Govindan R. Lung cancer in never smokers: a review. J Clin Oncol. 2007 Feb 10;25(5):561-70.

[22] Ferlay J, Shin HR, Bray F, Forman D, Mathers C and Parkin DM. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10.

[23] Cai H, Kumar N, Baudis M. arrayMap: a reference resource for genomic copy number imbalances in human malignancies. PLoS One. 2012;7(5):e36944.

[24] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D.M. Parkin et al. ICD-O-3. 2000; WHO, Geneva, Switzerland.

[25] National Cancer Institute. Cancer staging. 2013 Jan.

[26] Barrett T, Wilhite SE, Ledoux P, Evangelista C, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

[27] Rustici G, Kolesnikov N, Brandizi M, Burdett T, et al. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 2013 Jan; 41(Database issue):D987-90.

[28] Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. Genes Dev. 2011 Mar 15;25(6):534-55.

[29] Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011 Mar;17(3):297-303.

[30] Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 2013 Jan;41(Database issue):D64-9.

[31] Ou SH. Lung cancer in never-smokers. Does smoking history matter in the era of molecular diagnostics and targeted therapy? J Clin Pathol. 2013 May 9.

[32] Grewal P, Viswanathen VA. Liver cancer and alcohol. Clin Liver Dis. 2012 Nov;16(4): 839-50.

[33] Marrot L, Meunier JR. Skin DNA photodamage and its biological consequences. J Am Acad Dermatol. 2008 May;58(5 Suppl 2):S139-48.

[34] Bracht JR, Fang W, Goldman AD, Dolzhenko E, et al. Genomes on the edge: programmed genome instability in ciliates. Cell. 2013 Jan 31;152(3):406-16.

[35] Liu P, Carvalho CM, Hastings PJ, Lupski JR. Mechanisms for recurrent and complex human genomic rearrangements. Curr Opin Genet Dev. 2012 Jun;22(3):211-20.

[36] Chen JM, Cooper DN, Férec C, Kehrer-Sawatzki H, Patrinos GP. Genomic rearrangements in inherited disease and cancer. Semin Cancer Biol. 2010 Aug;20(4): 222-33.

[37] Malkova A, Haber JE. Mutations arising during repair of chromosome breaks. Annu Rev Genet. 2012;46:455-73.

[38] Bordeianu G, Zugun-Eloae F, Rusu MG. The role of DNA repair by homologous recombination in oncogenesis. Rev Med Chir Soc Med Nat Iasi. 2011 Oct-Dec; 115(4):1189-94.

[39] Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. 2009 Jan; 5(1):e1000327.

[40] Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN. Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. Hum Mutat. 2005 Oct;26(4):362-73.

[41] Zhang F, Carvalho CM, Lupski JR. Complex human chromosomal and genomic rearrangements. Trends Genet. 2009 Jul;25(7):298-307.

[42] Colnaghi R, Carpenter G, Volker M, O'Driscoll M. The consequences of structural genomic alterations in humans: genomic disorders, genomic instability and cancer. Semin Cell Dev Biol. 2011 Oct;22(8):875-85.

[43] Cheung KJ, Rogic S, Ben-Neriah S, Boyle M, et al. SNP analysis of minimally evolved t(14;18)(q32;q21)-positive follicular lymphomas reveals a common copy-neutral loss of heterozygosity pattern. Cytogenet Genome Res. 2012;136(1):38-43.

[44] González JR, Rodríguez-Santiago B, Cáceres A, Pique-Regi R, et al. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. BMC Bioinformatics. 2011 May 17;12:166.

[45] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004 Aug;10(8):789-99.

[46] Russnes HG, Vollan HK, Lingjaerde OC, Krasnitz A, et al. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. Sci Transl Med. 2010 Jun 30;2(38):38ra47.

[47] Stephens PJ, Tarpey PS, Davies H, Van Loo P, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012 May 16;486(7403):400-4.

[48] Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

[49] Yuan X, Yu G, Hou X, Shih IeM, et al. Genome-wide identification of significant aberrations in cancer genome. BMC Genomics. 2012 Jul 27;13:342.

[50] Beroukhim R, Getz G, Nghiemphu L, Barretina J, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007 Dec 11;104(50):20007-12. Epub 2007 Dec 6.

[51] Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput Biol. 2011 Oct;7(10):e1002227.

[52] Tang YC, Amon A. Gene copy-number alterations: a cost-benefit analysis. Cell. 2013 Jan 31;152(3):394-405.

[53] Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007 Apr;7(4):233-45. Epub 2007 Mar 15.

[54] Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008 Jul;8(7):497-511.

[55] Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer. 1997 Dec;20(4):399-407.

[56] Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet. 2009 Aug;85(2):142-54.

[57] Shortt J, Johnstone RW. Oncogenes in cell survival and cell death. Cold Spring Harb Perspect Biol. 2012 Dec 1;4(12). pii: a009829.

[58] Oxnard GR, Binder A, Jänne PA. New targetable oncogenes in non-small-cell lung cancer. J Clin Oncol. 2013 Mar 10;31(8):1097-104.

[59] Croce CM. Oncogenes and cancer. N Engl J Med. 2008 Jan 31;358(5):502-11.

[60] Todd R, Wong DT. Oncogenes. Anticancer Res. 1999 Nov-Dec;19(6A):4729-46.

[61] Cotter TG. Apoptosis and cancer: the genesis of a research field. Nat Rev Cancer. 2009 Jul;9(7):501-7.

[62] Zhang Y, Xia J, Zhang Y, Qin Y, et al. Pitfalls in experimental designs for characterizing the transcriptional, methylational and copy number changes of oncogenes and tumor suppressor genes. PLoS One. 2013;8(3):e58163.

[63] Forbes SA, Bindal N, Bamford S, Cole C, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011 Jan;39(Database issue):D945-50.

[64] Duffy MJ, Crown J. Companion Biomarkers: Paving the Pathway to Personalized Treatment for Cancer. Clin Chem. 2013 May 8.

[65] Dang CV. MYC on the path to cancer. Cell. 2012 Mar 30;149(1):22-35.

[66] Ohno S, Yazaki A. Simple construction of human c-myc gene implicated in B-cell neoplasmas and its relationship with avian v-myc and human lymphokines. Scand J Immunol. 1983 Nov;18(5):373-88.

[67] Dalla-Favera R, Bregni M, Erikson J, Patterson D, et al. Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proc Natl Acad Sci U S A. 1982 Dec;79(24):7824-7.

[68] Taub R, Kirsch I, Morton C, Lenoir G, et al. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc Natl Acad Sci U S A. 1982 Dec;79(24):7837-41.

[69] Mitchell KF, Battey J, Hollis GF, Moulding C, et al. The effect of translocations on the cellular myc gene in Burkitt lymphomas. J Cell Physiol Suppl. 1984;3:171-7.

[70] Paulson KG, Lemos BD, Feng B, Jaimes N, et al. Array-CGH reveals recurrent genomic changes in Merkel cell carcinoma including amplification of L-Myc. J Invest Dermatol. 2009 Jun;129(6):1547-55.

[71] Wu R, Lin L, Beer DG, Ellenson LH, et al. Amplification and overexpression of the L-MYC proto-oncogene in ovarian carcinomas. Am J Pathol. 2003 May;162(5):1603-10.

[72] Lv XH, Chen JW, Zhao G, Feng ZZ, et al. N-myc downstream-regulated gene 1/Cap43 may function as tumor suppressor in endometrial cancer. J Cancer Res Clin Oncol. 2012 Oct;138(10):1703-15. Epub 2012 Jun 8.

[73] Buechner J, Einvik C. N-myc and noncoding RNAs in neuroblastoma. Mol Cancer Res. 2012 Oct;10(10):1243-53.

[74] Lu X, Pearson A, Lunec J. The MYCN oncoprotein as a drug development target. Cancer Lett. 2003 Jul 18;197(1-2):125-30.

[75] Lin CY, Lovén J, Rahl PB, Paranal RM, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012 Sep 28;151(1):56-67.

[76] Liu H, Radisky DC, Yang D, Xu R, et al. MYC suppresses cancer metastasis by direct transcriptional silencing of αv and β3 integrin subunits. Nat Cell Biol. 2012 May 13;14(6):567-74.

[77] Brooks TA, Hurley LH. The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. Nat Rev Cancer. 2009 Dec;9(12):849-61.

[78] Radtke F, Raj K. The role of Notch in tumorigenesis: oncogene or tumour suppressor? Nat Rev Cancer. 2003 Oct;3(10):756-67.

[79] Delbridge AR, Valente LJ, Strasser A. The role of the apoptotic machinery in tumor suppression. Cold Spring Harb Perspect Biol. 2012 Nov 1;4(11). pii: a008789.

[80] Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971 Apr;68(4):820-3.

[81] MacPherson D, Dyer MA. Retinoblastoma: from the two-hit hypothesis to targeted chemotherapy. Cancer Res. 2007 Aug 15;67(16):7547-50.

[82] Berger AH, Knudson AG, Pandolfi PP. A continuum model for tumour suppression. Nature. 2011 Aug 10;476(7359):163-9.

[83] O'Keefe C, McDevitt MA, Maciejewski JP. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. Blood. 2010 Apr 8;115(14):2731-9.

[84] Thiagalingam S, Foy RL, Cheng KH, Lee HJ, et al. Loss of heterozygosity as a predictor to map tumor suppressor genes in cancer: molecular basis of its occurrence. Curr Opin Oncol. 2002 Jan;14(1):65-72.

[85] Jahromi MS, Putnam AR, Druzgal C, Wright J, et al. Molecular inversion probe analysis detects novel copy number alterations in Ewing sarcoma. Cancer Genet. 2012 Jul-Aug;205(7-8):391-404.

[86] Jasmine F, Rahaman R, Dodsworth C, Roy S, et al. A genome-wide study of cytogenetic changes in colorectal cancer using SNP microarrays: opportunities for future personalized treatment. PLoS One. 2012;7(2):e31968.

[87] Caron de Fromentel C, Soussi T. TP53 tumor suppressor gene: a model for investigating human mutagenesis. Genes Chromosomes Cancer. 1992 Jan;4(1): 1-15.

[88] Soussi T, Béroud C. Assessing TP53 status in human tumours to evaluate clinical outcome. Nat Rev Cancer. 2001 Dec;1(3):233-40.

[89] Maddocks OD, Berkers CR, Mason SM, Zheng L, et al. Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. Nature. 2013 Jan 24;493(7433):542-6.

[90] Junttila MR, Karnezis AN, Garcia D, Madriles F, et al. Selective activation of p53-mediated tumour suppression in high-grade tumours. Nature. 2010 Nov 25;468(7323):567-71.

[91] Laurie NA, Donovan SL, Shih CS, Zhang J, et al. Inactivation of the p53 pathway in retinoblastoma. Nature. 2006 Nov 2;444(7115):61-6.

[92] Vassilev LT, Vu BT, Graves B, Carvajal D, et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science. 2004 Feb 6;303(5659):844-8. Epub 2004 Jan 2.

[93] Speidel D. Transcription-independent p53 apoptosis: an alternative route to death. Trends Cell Biol. 2010 Jan;20(1):14-24.

[94] Palmero EI, Achatz MI, Ashton-Prolla P, Olivier M, Hainaut P. Tumor protein 53 mutations and inherited cancer: beyond Li-Fraumeni syndrome. Curr Opin Oncol. 2010 Jan;22(1):64-9.

[95] Srivastava S, Zou ZQ, Pirollo K, Blattner W, Chang EH. Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome. Nature. 1990 Dec 20-27;348(6303):747-9.

[96] Masciari S, Dillon DA, Rath M, Robson M, et al. Breast cancer phenotype in women with TP53 germline mutations: a Li-Fraumeni syndrome consortium effort. Breast Cancer Res Treat. 2012 Jun;133(3):1125-30.

[97] Muller PA, Vousden KH. p53 mutations in cancer. Nat Cell Biol. 2013 Jan;15(1):2-8.

[98] Beroukhim R, Mermel CH, Porter D, Wei G, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899-905.

[99] Grasso CS, Wu YM, Robinson DR, Cao X, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012 Jul 12;487(7406):239-43.

[100] Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010 Oct 28;467(7319):1109-13.

[101] Shen MM. Chromoplexy: a new category of complex rearrangements in the cancer genome. Cancer Cell. 2013 May 13;23(5):567-9.

[102] Roche B, Hochberg ME, Caulin AF, Maley CC, et al. Natural resistance to cancers: a Darwinian hypothesis to explain Peto's paradox. BMC Cancer. 2012 Sep 3;12:387.

[103] Torkamani A, Verkhivker G, Schork NJ. Cancer driver mutations in protein kinase genes. Cancer Lett. 2009 Aug 28;281(2):117-27.

[104] Bashashati A, Haffari G, Ding J, Ha G, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012 Dec 22;13(12):R124.

[105] Stephens PJ, Greenman CD, Fu B, Yang F, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011 Jan 7;144(1):27-40.

[106] Maher CA, Wilson RK. Chromothripsis and human disease: piecing together the shattering process. Cell. 2012 Jan 20;148(1-2):29-32.

[107] Meyerson M, Pellman D. Cancer genomes evolve by pulverizing single chromosomes. Cell. 2011 Jan 7;144(1):9-10.

[108] Stevens-Kroef M, Weghuis DO, Croockewit S, Derksen L, et al. High detection rate of clinically relevant genomic abnormalities in plasma cells enriched from patients with multiple myeloma. Genes Chromosomes Cancer. 2012 Nov;51(11):997-1006.

[109] Bass AJ, Lawrence MS, Brace LE, Ramos AH, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011 Sep 4;43(10):964-8.

[110] Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, et al. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol. 2011 Oct 19;12(10):R103.

[111] Zhang J, Ding L, Holmfeldt L, Wu G, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012 Jan 11;481(7380):157-63.

[112] Jiang Z, Jhunjhunwala S, Liu J, Haverty PM, et al. The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. 2012 Apr;22(4):593-601.

[113] Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. Nature. 2012 Feb 22;483(7391):589-93.

[114] Natrajan R, Mackay A, Lambros MB, Weigelt B, et al. A whole-genome massively parallel sequencing analysis of BRCA1 mutant oestrogen receptor-negative and -positive breast cancers. J Pathol. 2012 May;227(1):29-41.

[115] Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, et al. The life history of 21 breast cancers. Cell. 2012 May 25;149(5):994-1007.

[116] Northcott PA, Shih DJ, Peacock J, Garzia L, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012 Aug 2;488(7409): 49-56.

[117] Jones DT, Jäger N, Kool M, Zichner T, et al. Dissecting the genomic complexity underlying medulloblastoma. Nature. 2012 Aug 2;488(7409):100-5.

[118] Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet. 2011 May 15;20(10):1916-24.

[119] Chiang C, Jacobsen JC, Ernst C, Hanscom C, et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat Genet. 2012 Mar 4;44(4):390-7, S1.

[120] Deakin JE, Bender HS, Pearse AM, Rens W, et al. Genomic restructuring in the Tasmanian devil facial tumour: chromosome painting and gene mapping provide clues to evolution of a transmissible tumour. PLoS Genet. 2012;8(2):e1002483.

[121] Le LP, Nielsen GP, Rosenberg AE, Thomas D, et al. Recurrent chromosomal copy number alterations in sporadic chordomas. PLoS One. 2011;6(5):e18846.

[122] Magrangeas F, Avet-Loiseau H, Munshi NC, Minvielle S. Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. Blood. 2011 Jul 21;118(3):675-8.

[123] Kitada K, Taima A, Ogasawara K, Metsugi S, Aikawa S. Chromosome-specific segmentation revealed by structural analysis of individually isolated chromosomes. Genes Chromosomes Cancer. 2011 Apr;50(4):217-27.

[124] Poaty H, Coullin P, Peko JF, Dessen P, et al. Genome-wide high-resolution aCGH analysis of gestational choriocarcinomas. PLoS One. 2012;7(1):e29426.

[125] Rausch T, Jones DT, Zapatka M, Stütz AM, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012 Jan 20;148(1-2):59-71.

[126] Wu C, Wyatt AW, McPherson A, Lin D, et al. Poly-gene fusion transcripts and chromothripsis in prostate cancer. Genes Chromosomes Cancer. 2012 Dec;51(12): 1144-53.

[127] Lapuk AV, Wu C, Wyatt AW, McPherson A, et al. From sequence to molecular pathology, and a mechanism driving the neuroendocrine phenotype in prostate cancer. J Pathol. 2012 Jul;227(3):286-97.

[128] Berger MF, Hodis E, Heffernan TP, Deribe YL, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012 May 9;485(7399):502-6.

[129] Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. Cell Rep. 2012 Jun 28;1(6): 648-55.

[130] Govindan R, Ding L, Griffith M, Subramanian J, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. 2012 Sep 14;150(6): 1121-34.

[131] Kim TM, Xi R, Luquette LJ, Park RW, et al. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome Res. 2013 Feb;23(2):217-27.

[132] Zehentner BK, Hartmann L, Johnson KR, Stephenson CF, et al. Array-based karyotyping in plasma cell neoplasia after plasma cell enrichment increases detection of genomic aberrations. Am J Clin Pathol. 2012 Oct;138(4):579-89.

[133] Liu P, Erez A, Nagamani SC, Dhar SU, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell. 2011 Sep 16;146(6):889-903.

[134] Fullwood MJ, Lee J, Lin L, Li G, et al. Next-generation sequencing of apoptotic DNA breakpoints reveals association with actively transcribed genes and gene translocations. PLoS One. 2011;6(11):e26054.

[135] Tubio JM, Estivill X. Cancer: When catastrophe strikes a cell. Nature. 2011 Feb 24;470(7335):476-7.

[136] Crasta K, Ganem NJ, Dagher R, Lantermann AB, et al. DNA breaks and chromosome pulverization from errors in mitosis. Nature. 2012 Jan 18;482(7383): 53-8.

[137] Drets ME, Shaw MW. Specific banding patterns of human chromosomes. Proc Natl Acad Sci U S A. 1971 Sep;68(9):2073-7.

[138] Russell WO. Tissue diagnosis of cancer: changed concepts, technical advances and future applications. Proc Natl Cancer Conf. 1964;5:673-85.

[139] Bowen P. Chromosomal abnormalities. Clin Orthop Relat Res. 1964 Mar-Apr; 33:40-58.

[140] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8.

[141] Curtis C, Shah SP, Chin SF, Turashvili G, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Apr 18;486(7403):346-52.

[142] Langer-Safer PR, Levine M, Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci U S A. 1982 Jul;79(14): 4381-5.

[143] Cagir B, Gelmann A, Park J, Fava T, et al. Guanylyl cyclase C messenger RNA is a biomarker for recurrent stage II colorectal cancer. Ann Intern Med. 1999 Dec 7;131(11):805-12.

[144] Van Prooijen-Knegt AC, Van Hoek JF, Bauman JG, Van Duijn P, et al. In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp Cell Res. 1982 Oct; 141(2):397-407.

[145] Pinkel D, Segraves R, Sudar D, Clark S, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998 Oct;20(2):207-11.

[146] Zhao X, Li C, Paez JG, Chin K, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res. 2004 May 1;64(9):3060-71.

[147] Nannya Y, Sanada M, Nakazaki K, Hosoya N, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 2005 Jul 15;65(14):6071-9.

[148] Hall N. Advanced sequencing technologies and their wider impact in microbiology. J Exp Biol. 2007 May;210(Pt 9):1518-25.

[149] Church GM. Genomes for all. Sci Am. 2006 Jan;294(1):46-54.

[150] National Academy of Sciences. Science. 1960 Nov 18;132(3438):1488-501.

[151] Bobrow M, Madan K. The effects of various banding procedures on human chromosomes, studied with acridine orange. Cytogenet Cell Genet. 1973;12(3): 143-56.

[152] Caspersson T, Zech L, Johansson C. Differential binding of alkylating fluorochromes in human chromosomes. Exp Cell Res. 1970 Jun;60(3):315-9.

[153] Rowley JD, Bodmer WF. Relationship of centromeric heterochromatin to fluorescent banding patterns of metaphase chromosomes in the mouse. Nature. 1971 Jun 25;231(5304):503-6.

[154] Crossen PE. Giemsa banding patterns of human chromosomes. Clin Genet. 1972;3(3):169-79.

[155] Bühler EM, Tsuchimoto T, Stalder GR. Reverse banding in the human Y chromosome. Lancet. 1973 May 26;1(7813):1178-9.

[156] Speicher MR, Carter NP. The new cytogenetics: blurring the boundaries with molecular biology. Nat Rev Genet. 2005 Oct;6(10):782-92.

[157] Daniel A, Lam-Po-Tang PR. Mechanism for the chromosome banding phenomenon. Nature. 1973 Aug 10;244(5415):358-9.

[158] Northcott PA, Rutka JT, Taylor MD. Genomics of medulloblastoma: from Giemsa-banding to next-generation sequencing in 20 years. Neurosurg Focus. 2010 Jan; 28(1):E6.

[159] Cavazzini F, Ciccone M, Negrini M, Rigolin GM, Cuneo A. Clinicobiologic importance of cytogenetic lesions in chronic lymphocytic leukemia. Expert Rev Hematol. 2009 Jun;2(3):305-14.

[160] O'Connor M, Peifer M, Bender W. Construction of large DNA segments in Escherichia coli. Science. 1989 Jun 16;244(4910):1307-12.

[161] Shizuya H, Birren B, Kim UJ, Mancino V, et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A. 1992 Sep 15;89(18):8794-7.

[162] Shizuya H, Kouros-Mehr H. The development and applications of the bacterial artificial chromosome cloning system. Keio J Med. 2001 Mar;50(1):26-30.

[163] Burke DT, Carle GF, Olson MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science. 1987 May 15;236(4803): 806-12.

[164] Weiss MM, Hermsen MA, Meijer GA, van Grieken NC, et al. Comparative genomic hybridisation. Mol Pathol. 1999 Oct;52(5):243-51.

[165] Ren H, Francis W, Boys A, Chueh AC, et al. BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints. Hum Mutat. 2005 May;25(5):476-82.

[166] Bignell GR, Huang J, Greshock J, Watt S, et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res. 2004 Feb;14(2): 287-95.

[167] Lockwood WW, Chari R, Chi B, Lam WL. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. Eur J Hum Genet. 2006 Feb;14(2):139-48.

[168] Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res. 2006 Jan 26;34(2):445-50. Print 2006.

[169] Baudis M, Cleary ML. Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics. 2001 Dec;17(12):1228-9.

[170] Cowell JK, Hawthorn L. The application of microarray technology to the analysis of the cancer genome. Curr Mol Med. 2007 Feb;7(1):103-20.

[171] Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, et al. A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet. 2004 Mar; 36(3):299-303. Epub 2004 Feb 15.

[172] Heiskanen MA, Bittner ML, Chen Y, Khan J, Adler KE, Trent JM, Meltzer PS. Detection of gene amplification by genomic hybridization to cDNA microarrays. Cancer Res. 2000 Feb 15;60(4):799-802.

[173] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. 1999 Sep;23(1):41-6.

[174] Brennan C, Zhang Y, Leo C, Feng B, et al. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. Cancer Res. 2004 Jul 15;64(14): 4744-8.

[175] Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. J Clin Pathol. 2004 Jun;57(6):644-6.

[176] Urban AE, Korbel JO, Selzer R, Richmond T, et al. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci U S A. 2006 Mar 21;103(12):4534-9. Epub 2006 Mar 14.

[177] Shinawi M, Cheung SW. The array CGH and its clinical applications. Drug Discov Today. 2008 Sep;13(17-18):760-70.

[178] Lucito R, Healy J, Alexander J, Reiner A, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res. 2003 Oct;13(10):2291-305. Epub 2003 Sep 15.

[179] Le Scouarnec S, Gribble SM. Characterising chromosome rearrangements: recent technical advances in molecular cytogenetics. Heredity (Edinb). 2012 Jan;108(1): 75-85.

[180] Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet. 2005 Jun;37 Suppl:S11-7.

[181] Gardina PJ, Lo KC, Lee W, Cowell JK, Turpaz Y. Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. BMC Genomics. 2008 Oct 17;9:489.

[182] Lo KC, Bailey D, Burkhardt T, Gardina P, et al. Comprehensive analysis of loss of heterozygosity events in glioblastoma using the 100K SNP mapping arrays and comparison with copy number abnormalities defined by BAC array comparative genomic hybridization. Genes Chromosomes Cancer. 2008 Mar;47(3):221-37.

[183] Schaaf CP, Wiszniewska J, Beaudet AL. Copy number and SNP arrays in clinical diagnostics. Annu Rev Genomics Hum Genet. 2011 Sep 22;12:25-51.

[184] Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet. 2004 Dec;66(6): 488-95.

[185] Brady PD, Vermeesch JR. Genomic microarrays: a technology overview. Prenat Diagn. 2012 Apr;32(4):336-43.

[186] International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299-320.

[187] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851-61.

[188] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55.

[189] Yau C, Holmes CC. CNV discovery using SNP genotyping arrays. Cytogenet Genome Res. 2008;123(1-4):307-12.

[190] Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics. 2005 Oct 1;21(19):3763-70. Epub 2005 Aug 4.

[191] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004 Oct;5(4): 557-72.

[192] Cao Q, Zhou M, Wang X, Meyer CA, et al. CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic Acids Res. 2011 Jan;39(Database issue):D968-74.

[193] Scheinin I, Myllykangas S, Borze I, Böhling T, et al. CanGEM: mining gene copy number changes in cancer. Nucleic Acids Res. 2008 Jan;36(Database issue):D830-5. Epub 2007 Oct 11.

[194] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931-45.

[195] Kulldorff, M. A spatial scan statistic. Commun statist. 1997;26(6):1481–1496.

[196] Bengtsson, H., Simpson, K., Bullard, J., and Hansen, K. aroma.affymetrix: A genetic framework in R for analyzing small to very large affymetrix data sets in bounded memory. 2008; Tech Report #745, Department of Statistics, University of California, Berkeley.

[197] Thieme S, Groth P. Genome Fusion Detection: a novel method to detect fusion genes from SNP-array data. Bioinformatics. 2013 Mar 15;29(6):671-7.

[198] Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer. 2012 Oct;12(10): 663-70.

[199] Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Res. 2008 Jan;36(Database issue):D440-4. Epub 2007 Nov 4.